2012

Face analysis based on reference samples

Ce Zhan

University of Wollongong

Recommended Citation
UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
Face Analysis Based on Reference Samples

A thesis submitted in fulfillment of the requirements for the award of the degree

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

by

Ce Zhan

School of Computer Science and Software Engineering
November 2012
© Copyright 2012

by

Ce Zhan

All Rights Reserved
Dedicated to

My grandparents, Hongchao Li and Junhui Dong
Declaration

This is to certify that the work reported in this thesis was done by the author, unless specified otherwise, and that no part of it has been submitted in a thesis to any other university or similar institution.

Ce Zhan
November 17, 2012
Abstract

Humans are not sensitive to variations in facial appearance and are capable of performing face analysis tasks reliably under realistic conditions when compared with current computer-based face analysis techniques. This can be partly explained by the ability of humans to make effective use of previously encountered known faces for both internal representation and processing.

This thesis focuses on establishing computational models to account for the cognitive findings related to internal face representation and two fundamental perception processes (distinctiveness and familiarity), and developing novel methods based on the models for face analysis. Specifically, a set of reference samples that may or may not contain any labeling information and any instance of the person whose face is under consideration are proposed to model previously encountered faces. The non-negative matrix factorization which affords part-based representation is extended to learn reusable local facial patterns for representation from the reference set. Computational models are developed for locating distinctive areas and measuring familiarity of faces with respect to the reference set. By employing the proposed face representation, distinctiveness and familiarity models, novel schemes are developed to recognize faces from single sample per person and estimate ages and head poses of faces.
Acknowledgments

One of the joys of completion is to look over the journey and remember all the people who have helped and supported me along this long but fulfilling road. I would like to take this opportunity to express my sincere gratitude to my supervisors, A/Prof. Wanqing Li and Prof. Philip Ogunbona. This work would not have been possible without their invaluable guidance, advice, criticism and encouragement. Under their guidance I learned a lot and successfully overcame so many difficulties. For me, they are more of role models and life mentors than professors.

I am very grateful to my friends Jun Hu, Aimee Zhang, Yang Yang, Xiaodong Yang and Yiyu Qiu, they took care of me just like elder brothers and sisters. I also wish to thank my colleague Hongda Tian and Duc Nguyen for numerous discussions which have given me tremendous confidence and inspiration. I must also acknowledge Nikola Milosevic and Michael Milway for proofreading my papers and thesis chapters.

I would like to express my eternal gratitude to my parents for their everlasting love, encouragement and support. In particular, the patience and understanding shown by them during these years are greatly appreciated.

I am deeply and forever indebted to my grandparents Hongchao Li and Junhui Dong for their unconditional love throughout my entire life. It is to them that I dedicate this work. All the efforts I made are to make them proud.
Contents

Abstract v

Acknowledgments vi

1 Introduction 1

1.1 Motivation 1

1.2 Contributions 3

1.3 Publication list 4

1.3.1 Published papers 4

1.3.2 Papers under revision and to be submitted 5

1.4 Organization of thesis 5

2 Literature review 7

2.1 Psychological aspects of human face perception 7

2.1.1 Internal representation 7

2.1.2 Distinctiveness and familiarity 9

2.2 Single sample-based face recognition 12

2.2.1 Face to face matching 15

2.2.2 Use of virtual samples 19

2.2.3 Use of the training samples of other subjects 21

2.2.4 Use of generic samples 24

2.3 Age estimation 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 Geometric-based approach</td>
<td>27</td>
</tr>
<tr>
<td>2.3.2 Appearance-based approach</td>
<td>27</td>
</tr>
<tr>
<td>2.4 Head pose estimation</td>
<td>28</td>
</tr>
<tr>
<td>2.4.1 Geometric-based approach</td>
<td>29</td>
</tr>
<tr>
<td>2.4.2 Appearance-based approach</td>
<td>30</td>
</tr>
<tr>
<td>2.5 Databases for face analysis</td>
<td>31</td>
</tr>
<tr>
<td>2.5.1 Databases for face recognition</td>
<td>32</td>
</tr>
<tr>
<td>2.5.2 Databases for age estimation</td>
<td>34</td>
</tr>
<tr>
<td>2.5.3 Databases for head pose estimation</td>
<td>36</td>
</tr>
<tr>
<td>2.6 Summary</td>
<td>36</td>
</tr>
<tr>
<td>3 Localized compact face representation</td>
<td>38</td>
</tr>
<tr>
<td>3.1 Non-negative matrix factorization</td>
<td>39</td>
</tr>
<tr>
<td>3.2 Extended NMF</td>
<td>41</td>
</tr>
<tr>
<td>3.3 Experimental results</td>
<td>45</td>
</tr>
<tr>
<td>3.3.1 Visual comparison</td>
<td>46</td>
</tr>
<tr>
<td>3.3.2 Objective evaluation</td>
<td>46</td>
</tr>
<tr>
<td>3.3.3 Computational complexity</td>
<td>50</td>
</tr>
<tr>
<td>3.3.4 The discriminative power of ENMF</td>
<td>51</td>
</tr>
<tr>
<td>3.4 Summary</td>
<td>53</td>
</tr>
<tr>
<td>4 Modeling distinctiveness and familiarity</td>
<td>54</td>
</tr>
<tr>
<td>4.1 Finding distinctive facial regions</td>
<td>54</td>
</tr>
<tr>
<td>4.1.1 Construction of a general face model</td>
<td>55</td>
</tr>
<tr>
<td>4.1.2 Measurement of distinctiveness</td>
<td>55</td>
</tr>
<tr>
<td>4.1.3 Experimental results</td>
<td>56</td>
</tr>
<tr>
<td>4.2 Measuring face familiarity</td>
<td>60</td>
</tr>
<tr>
<td>4.2.1 The proposed method</td>
<td>60</td>
</tr>
<tr>
<td>4.2.2 Experimental results</td>
<td>64</td>
</tr>
<tr>
<td>4.3 Summary</td>
<td>70</td>
</tr>
</tbody>
</table>
8 Conclusion

8.1 Summary ....................................................... 118
8.2 Future works .................................................. 119

Bibliography ....................................................... 121
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>LC for different representations learned from FERET database</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>LC for different representations learned from ORL database</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Typical CPU times, factorization approximation errors, and number of iterations</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>The recognition rates for different methods</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>The recognition rates for different methods in single sample scenario</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>The percentages of each type of reference groups selected as the most, second most, third most similar group to the testing faces</td>
<td>70</td>
</tr>
<tr>
<td>5.1</td>
<td>The recognition rates of different methods for the standard FERET evaluation protocols</td>
<td>80</td>
</tr>
<tr>
<td>5.2</td>
<td>The recognition rates of different methods for the standard FERET evaluation protocols</td>
<td>92</td>
</tr>
<tr>
<td>6.1</td>
<td>The recognition rates of different methods based on the standard FERET evaluation protocols</td>
<td>96</td>
</tr>
<tr>
<td>6.2</td>
<td>The recognition rates of different methods for the standard FERET evaluation protocols</td>
<td>104</td>
</tr>
<tr>
<td>6.3</td>
<td>The verification accuracies of different methods based on the image-restricted training protocol in LFW database view 2</td>
<td>106</td>
</tr>
<tr>
<td>7.1</td>
<td>The average accuracies of subspace-based head pose estimation methods</td>
<td>111</td>
</tr>
<tr>
<td>7.2</td>
<td>Age group distribution of the images in MORPH and FG-NET database</td>
<td>112</td>
</tr>
</tbody>
</table>
7.3 The classification accuracy for 4 age groups .......................... 114
7.4 The classification accuracies for 2 age groups ......................... 114
7.5 The classification accuracy for 14 age group estimations .......... 115
7.6 The MAE for exact age estimation, ENMF(nb) denotes the proposed method without using border groups ................................. 117
List of Figures

2.1 Thatcher illusion: A face can be “Thatcherized” by inverting the eyes and the mouth, and then inverting the entire picture. Most people do not notice anything peculiar about the inverted face. Upright, however, we see a gross distortion of the configuration of the facial parts [146].

2.2 An illustration of a face LEM [32]

2.3 An example of the Face-ARG representation and partial matching [115]

2.4 An example of the projection map and projection combined image in [163]

2.5 An example of the first-order and second-order projection map and projection combined image in [17]

2.6 Sample images from the AR face database

2.7 Sample images from ORL Face database

2.8 Sample images from BioID Face database

2.9 Sample images from BioID Face database

2.10 Sample images from LFW Face database

2.11 Sample images of the five frontal image subsets of FERET database

2.12 Sample images from FG-NET Face database

2.13 Sample images from MORPH Face database

2.14 Sample images from CMU PIE face database

3.1 Basis images learned from ORL database using different methods ($r = 49$).

3.2 ENMF basis images learnt from normalized ORL database with $r = 81$, a group of ENMF bases related to the “mouth” area are marked with red squares.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Sample face images used in the experiments</td>
<td>49</td>
</tr>
<tr>
<td>3.4</td>
<td>Typical objective function vs. the number of iterations for NMF, LNMF, NMFsc and ENMF</td>
<td>51</td>
</tr>
<tr>
<td>3.5</td>
<td>The CPU times of ENMF learning with different values of ψ and n</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>The normalization process</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>Examples of normalized faces from BioID database</td>
<td>57</td>
</tr>
<tr>
<td>4.3</td>
<td>The ENMF bases learned from the BioID dataset</td>
<td>57</td>
</tr>
<tr>
<td>4.4</td>
<td>Examples of original face image (bottom row) and the corresponding reconstructed image (top row) in the learned subspace</td>
<td>58</td>
</tr>
<tr>
<td>4.5</td>
<td>Results for exaggerated faces</td>
<td>58</td>
</tr>
<tr>
<td>4.6</td>
<td>An sample face image with the centers (“x”) of distinctive areas labeled by the participants, the mean center (“+”) and the most distinctive 16×16 region (the red rectangle) detected by the proposed algorithm</td>
<td>59</td>
</tr>
<tr>
<td>4.7</td>
<td>The concept of the proposed familiarity measure</td>
<td>62</td>
</tr>
<tr>
<td>4.8</td>
<td>Examples of morph (the second row) and their parent faces (the first and third rows)</td>
<td>65</td>
</tr>
<tr>
<td>4.9</td>
<td>The familiarity of three sample faces (κ = 10, ζ = 10 using personal reference groups) and the faces of their ten most similar subjects in the reference set (left to right, from the most similar to the least)</td>
<td>66</td>
</tr>
<tr>
<td>4.10</td>
<td>Results for using clustered reference groups in the case of C = 1</td>
<td>67</td>
</tr>
<tr>
<td>4.11</td>
<td>Results for using clustered reference groups in the case: C = 10, κ = 3</td>
<td>67</td>
</tr>
<tr>
<td>4.12</td>
<td>Results for using personal reference groups when ζ = 10</td>
<td>68</td>
</tr>
<tr>
<td>4.13</td>
<td>Results for using clustered reference groups with C = 50, ζ = 10</td>
<td>69</td>
</tr>
<tr>
<td>4.14</td>
<td>A snapshot of the training interface in subjective study</td>
<td>71</td>
</tr>
<tr>
<td>4.15</td>
<td>A snapshot of the testing interface in subjective study</td>
<td>71</td>
</tr>
<tr>
<td>4.16</td>
<td>Results for using combined reference groups with κ = 3</td>
<td>71</td>
</tr>
<tr>
<td>4.17</td>
<td>Results of the subjective ranking with error bars</td>
<td>71</td>
</tr>
<tr>
<td>5.1</td>
<td>Overall framework of local-based methods</td>
<td>74</td>
</tr>
</tbody>
</table>
5.2 Different local facial region partitions: a. regular grids with fixed size, b. regions around central salient facial points

5.3 Facial region partition (a) and weights (b) for the local regions in [2]. The facial image is equally divided into non-overlapped 7×7 blocks. The black squares indicate weight 0.0, dark gray 1.0, light gray 2.0 and white 4.0

5.4 The block diagram of improved local-based methods inspired by distinctiveness measure

5.5 The block diagram of the proposed hierarchical method

5.6 A hierarchically partitioned face image. The face is divided into 6 regions $G_i$ (red blocks), each $G_i$ is further divided into sub-regions $g_{k,i}$ (blue blocks).

5.7 Examples of original face image and the corresponding reconstructed images (DCT features were used)

5.8 Typical examples of regional histograms of left eye region (DCT features were used)

5.9 Typical examples of regional histograms of left eye region (ENMF features were used)

5.10 Effect of different DCT subsets

5.11 Effect of different sub-block overlaps (DCT feature)

5.12 Effect of different codebook sizes (DCT feature)

5.13 Effect of different codebook sizes (LBP feature)

5.14 Effect of different sub-block overlaps (LBP feature)

6.1 The familiarity vectors of three face images

6.2 Face recognition using a NN classifier in a familiarity space

6.3 The conceptual diagram of the proposed method for matching two faces in a familiarity space

6.4 Normalized images of one person from AR database with index of 1-13

6.5 Normalized images of one person from AR database with index of 14-26
7.1 Sample face images of one person from CMU PIE database; the numbers beneath each image represent the poses ................................. 109
7.2 The testing results for different value of $S_h$ while $r$ is fixed to 49 ........... 110
7.3 The testing results for different value of $r$ while $S_h$ is fixed to 0.3 ........... 111
7.4 Comparison of the estimation accuracy for each pose by different methods . 111
7.5 The classification accuracy for different values of $S_h$ while $r = 81$ ........... 113
7.6 The classification accuracy for different values of $r$ while $S_h = 0.3$ ........... 113
7.7 The coarse to fine scheme for exact age estimation ............................... 116
1.1 Motivation

The human face conveys rich information about race, gender, identity and age. In addition to these overt cues, the internal emotional status of an individual can be expressed indirectly through facial expressions. Face analysis is the process of extracting any of the information by analyzing the visual appearance of a person as recorded in digital images. Typical face analysis tasks include face recognition [186] [79] [105], facial expression recognition [114] [28] [187], gender classification [64] [6], and age estimation [36] [31]. In many applications, face analysis often acts as one of the key components in a system. For instance, an intelligent visual surveillance application comprises several functional modules and will use face analysis tools to automate functions ranging from simply counting people in a scene of interest to complex identification of people in a scene and analyzing their behavior. In human-computer interaction (HCI), face analysis enables the computer to recognize users and understand their intention and emotion thereby achieving more friendly and realistic interactions. In security, face recognition has been one of the most popular biometric modalities and is being widely adopted for national border control and computer access authentication (login). With the prevalence of on-line visual contents, face analysis can potentially lead to effective tools for automatic annotation, management and retrieval of the content by providing identity, race, gender, ages and other information of the people appearing in the images and video.

Driven by numerous potential applications, the past few decades have witnessed sustained research activities in face analysis by the computer vision, pattern recognition and
machine learning research community. Since the first face recognition system was reported by Kanade [58] almost 40 years ago, face analysis techniques have shifted gradually from a signal processing paradigm towards the contemporary machine learning approaches. Consequently, the significant progress made in automatic face analysis, as evidenced by recent literature surveys [186, 79, 105, 114, 28, 187, 64, 6, 36, 31], is mainly due to advances in the theories and tools of machine learning. Despite such advances, the performance of many state-of-the-art face analysis methods is still incomparable to the performance that an average person can achieve especially when the faces are acquired under realistic and unconstrained environment. Such faces often bear a wide range of variations in pose, illumination, expression and occlusion.

One of the reasons adduced for the poor performance, compared to humans, of existing face analysis methods is the insufficient understanding and modeling of the visual processes involved in human perception of faces. Study by psychologists and cognitive scientists has found that face perception involved a complex process and many factors. Recently, researchers have attempted to explicitly and implicitly mimic various aspects of human perception of faces in order to achieve robust and improved performance. For instance, Sinha et al. [135] provided and discussed 19 observations on how humans recognize faces as useful tools for computer vision researchers. In a similar vein, Schwaninger et al. [131] proposed a computational framework of face perception based on the premise that human face processing uses both “featural” and “configural” information; and Tistarellia et al. [147] mimicked the dynamic process of human visual perception for face recognition.

However, one of the important cognitive findings related to human perception of faces has not been well exploited in face analysis. In this finding it is argued that humans in fact make effective use of previously encountered faces in forming internal face representation and driving the process of face perception. Cognitive study [155] has suggested that the internal representation of faces in humans is very likely based on the combination of reusable local facial patterns that are abstracted or derived from previously experienced faces. This representation effectively encodes the knowledge about faces and is employed in the process of learning and understanding faces. In addition, when a new face is encountered, humans
process the face in a comparative manner [150]; the new face is compared with previously seen ones. Depending on the task of the processing, the previously seen faces are organized accordingly to accommodate the comparison to reveal the characteristics of the new face. In particular, there are two fundamental processes involved in the comparison. One is to reveal and locate the features or parts of the new face that are distinctive from the existing ones, i.e. distinctiveness. The other process is to reveal how and in what way the new face is (or becomes) familiar based on previously seen faces, i.e. familiarity. Study has found that the process for identifying the distinctiveness of a face is different from the process of measuring familiarity of the face and getting familiar with the face. In the former, a general face model abstracted from the previously seen faces is often employed, whereas, in the latter, familiarity is affected by the exposure frequency, exposure intensity and similar exposure to the previously seen faces. Nevertheless, the processes of distinctiveness and familiarity usually collaborate with each other in face perception. However, they may play different roles depending on the tasks of face analysis. For example, cognitive study [43] have shown that faces with distinctive facial features are easy to remember, and caricatures which exaggerate the distinctive facial parts are recognized more accurately than veridical versions of the faces. Compared to unfamiliar faces, humans respond to familiar faces more accurately and quickly in identification [121], gender classification [126] and age estimation [14].

1.2 Contributions

This thesis focuses on establishing computational models to account for the cognitive finding presented above and developing novel methods based upon the models for face analysis, especially, face recognition from single samples, age estimation and head pose estimation. Specifically,

- Reference face samples, referred to as a reference set, are proposed to represent the previously encountered faces. The reference set is essentially different from the traditional positive and negative training samples used in machine learning. It can be
arbitrary and may or may not contain instances of the same subject of the given face under consideration. In addition, face samples in the reference set are not necessarily labeled.

- The non-negative matrix factorization is extended to learn an effective localized compact representation of the “previously experienced” facial patterns from the reference set.

- A computational method is developed to quantitatively measure the distinctiveness of facial areas with respect to a general face model learned from the reference set.

- A computational model is proposed to objectively measure the degree of familiarity for a given face with respect to the reference set.

- Based on the compact representation and computational models for distinctiveness and familiarity, novel schemes are proposed to recognize faces from single samples.

- A new classifier, that can be considered as a special case of the proposed familiarity measure, is devised and employed in the tasks of pose and age estimation.

### 1.3 Publication list

#### 1.3.1 Published papers


1.3.2 Papers under revision and to be submitted


2. Ce Zhan, Wanqing Li, and Philip Ogunbona. Single Sample Face Recognition based on Reference Samples. (in preparation, to be submitted to *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*)

1.4 Organization of thesis

After the introduction, the rest of this thesis is organized as follows:

- **Chapter 2** presents the literature review that is organized into three parts. Firstly, the key psychological findings about human face perception are summarized. Secondly, existing methods for face analysis considered in this thesis (single sample face
recognition, age estimation and head pose estimation) are reviewed. Finally, the major publicly available databases for face analysis are briefly described.

- **Chapter 3** presents the proposed extended non-negative matrix factorization (ENMF) for realizing a local compact face representation. After a brief discussion on the traditional NMF and its major variants, details of the ENMF algorithm is described. The efficiency of ENMF is then evaluated based on benchmark databases and a novel objective measure of locality and compactness of a representation.

- **Chapter 4** describes the proposed computational models for quantitative measures of distinctiveness and familiarity. The methods are evaluated through experiments that follow the protocols commonly used in psychological study and compared with subjective evaluation.

- **Chapter 5** and **Chapter 6** present the single sample-based face recognition methods that are built upon the ENMF representation and computational models for distinctiveness and familiarity. In **Chapter 5**, the distinctiveness model is used to adaptively weight (select) facial regions for local-based methods and integrated with the configural information of faces through a hierarchical method. **Chapter 6** introduces the concept of familiarity spaces and presents methods for matching two faces in a familiarity space.

- **Chapter 7** extends the familiarity measure by utilizing the labelling information of the reference set and presents two methods for head pose and age estimation respectively.

- **Chapter 8** concludes the thesis and suggests possible future works.
Chapter 2

Literature review

This chapter provides a critical review of the literature related to the topics covered in this thesis. Specifically, the key findings in the recent cognitive studies about human perception of faces are first summarized and discussed. These findings lay the fundamental basis of the thesis. We then review the theories and methods related to the focus of the thesis, namely, single sample face recognition, age estimation and head pose estimation. As an introduction to the evaluation methods used in the thesis, a summary of the commonly used benchmarking datasets for development and evaluation of face analysis algorithms are presented.

2.1 Psychological aspects of human face perception

2.1.1 Internal representation

Within the psychological research literature of face perception, there has been considerable debate as to whether faces are represented holistically or locally in the human mind. Researchers who support the holistic model suggest that faces are encoded and processed as a whole entity without representing nameable facial parts (eyes, nose, mouth, etc.) explicitly [144, 27, 48]. They argue that discrimination based on one facial part is disrupted by the presence of other irrelevant facial parts. Even if attention is focused on only one part, the holistic context still contributes to the processing of individual features.

An evidence supporting the holistic model is that human face recognition is sensitive to the layout and the spacing of the nameable facial parts. Changes in the distances between facial parts such as the eyes, nose and mouth can greatly affect the recognition performance
2.1. Psychological aspects of human face perception

Figure 2.1: Thatcher illusion: A face can be “Thatcherized” by inverting the eyes and the mouth, and then inverting the entire picture. Most people do not notice anything peculiar about the inverted face. Upright, however, we see a gross distortion of the configuration of the facial parts [146].

[73]. This observation is referred to as configural effect in the psychological literature. An example illustrating the configural effect is the difficulty of recognizing an inverted face; the configuration information is disrupted by inversion. Another widely used example is the classic “Thatcher illusion” [146] (see Figure 2.1), which inverts the eyes and mouth and then invert the entire face image. The sensitivity to facial configuration has led some theorists to argue that humans represent faces using a code based on facial metrics. That is to say we use distances between landmarks such as the eye centers, tip of the nose, etc.

The holistic model is challenged by many researchers who argue that explicit part-based featural information is used in human face processing. The work of Schwaninger et al. [130] provided a good illustration for this argument. In their work, experiments show that both familiar and unfamiliar faces could be recognized by humans when they were scrambled into constituent parts. Recently, Sadr et al. [128] have shown that even just one facial feature (such as the eyes or, notably, the eyebrows) could be enough for the recognition of many famous faces.

Although it is easy to argue that humans process faces by encoding and storing both holistic configural and local featural information, incorporating all the pieces of evidence into a single model of human face representation is a difficult task. Related attempts have often led to devising hybrid systems [97, 118]. Recently, Wallis et al. [155] managed to
explain both the effects of configural and featural human face processing based on “competitive networks of neurons selective for environment specific abstract features”. Generally speaking, they suggest that faces in human mind are represented as combinations of reusable, abstract features that are derived from previously experienced local facial patterns. The abstract features are essentially pictorial sub-elements of a stimulus that display some degree of robustness to natural changes in the appearance of their preferred stimuli. This robustness may take the form of tolerance to changes in appearance of that feature caused by changes in object size, location, orientation in depth, illumination, etc. The abstract features not only encode nameable facial parts, but also encode the facial areas that span the nameable parts and carry the configural information. Therefore, the abstract feature model could respond to both local and holistic facial changes and the featural and configural face processing are integrated.

2.1.2 Distinctiveness and familiarity

Although the exact process of human face perception is still unknown, some psychologists have already concluded that the process is in a comparative manner [150, 109], where the face being perceived is compared with the previously seen ones. Two fundamental processes are involved during the comparison and are related to the generation of the feelings of distinctiveness and familiarity.

Unlike many other classes of objects, faces share the same overall configuration of parts on a similar scale. The convolutions in depth of the face surface vary little from one face to another, and for most parts, the reflectance distributions across different faces are also very similar [136]. It seems that the high overall similarity of faces should make face perception, especially face recognition, a very difficult task. However, humans can recognize faces quickly and accurately. Psychologists have suggested this is mainly due to our ability for identifying and encoding the distinctive facial parts/features [81]. Related studies indicated that humans tend to store a model of an average or general face that is abstracted from the previously seen faces [150]. This general face model that is supposedly stored in the mind gives us the internalized knowledge about faces such as how long a nose should be or how
close together two eyes should be. By comparing a newly encountered face against the gen-
eral face model in mind, humans quickly perceive the distinctiveness of the face and have
such feelings as “the nose is too big” or “the eyes are too close together”, etc. Each human
face is unique, however. The unique facial features for some faces are easy to identify while
for others, they are not. As a result, faces with distinct unique features are easy to remem-
ber and are usually recognized accurately. The recognition of caricatures provides a good
illustration of the above arguments on distinctiveness. Several studies have shown that car-
icatures which exaggerate the distinctive features in a face are recognized more accurately
and more quickly than veridical versions of the faces. This is regardless of the caricatures
being drawn by an artist or generated using a computer by exaggerating features that deviate
substantially from the average face [124].

Besides distinctiveness, humans also naturally generate the feeling of familiarity when
a face is encountered. Psychological studies have found that familiarity is one of the key
factors that affect human face processing. For instance, people are excellent at recognizing
faces that are familiar to them, even from very low quality images across large pose and
illumination variations. However, our ability to recognize unfamiliar face is rather poor
[121]. In addition, Humans respond to familiar faces more accurately and quickly in gender

In cognitive studies, face familiarity is defined qualitatively based on previous exposures.
Faces belonging to the people who have been seen many times or for a long duration are con-
sidered as familiar faces [121], whereas faces that have never been seen before or have been
seen for only a limited number of times or a short duration are considered as unfamiliar
faces [121]. The cognitive studies conducted so far on familiarity [177] have revealed three
major factors that affect human perception of face familiarity: exposure frequency, exposure
intensity, and similar exposure. Firstly, the feeling of familiarity increases after each stim-
ulus; the more times we see a face, the more familiar the face becomes to us. Secondly,
how we perceive a face each time will also impact the degree of familiarity. Seeing a face
in different conditions such as various views and lightings leads to a substantial increase
in familiarity than seeing the face in the same or similar conditions. Thirdly, prior exposure to the faces that belong to different persons also contributes to the familiarity of faces of unknown persons. For example, we often feel that a particular face looks familiar, yet the person has never been met before. He/she only bears a resemblance to several acquaintances. Psychologists refer to such phenomenon as the prototype effect [137]. Experiments of old/new discrimination and familiarity ranking tasks have verified that the prototype effect is prominent in face perception. The prototype faces that are generated from previously seen faces are not only regarded as familiar, but even more familiar than the previously seen faces [155].

The general concept of familiarity described by Mandler [92] may somehow account for the prototype effect. Mandler suggested two forms of familiarity: context-free familiarity and context-dependent familiarity. The former is a sense of knowing that we have encountered the target before, but no specific source in the memory contributes to this feeling of knowing. The latter is a feeling of remembering which is engendered when the target is matched to previously encountered faces in the memory. The prototype faces tend to cause strong feeling of context-free familiarity. However as Mandler suggested, we are not able to distinguish between these two forms of familiarity.

Wallis et al. [155] successfully explain the prototype effect from representation point of view based on the proposed reusable abstract features. They argue that holistic representation of the known faces will only rarely be fooled by a facial prototype and will never be thought to be less familiar than a prototype. This is because at least some global aspects of the prototype will always be novel and, hence, different from the faces seen previously. On the other hand, when the representation is based on certain abstracted patterns of unrelated small facial regions, the local features of prototypes are highly likely to match those in the memory and thus explain the prototype effect.
2.2 Single sample-based face recognition

The problem of single sample-based face recognition is to identify and recognize from a face image a person for whom only one sample face image is available for comparison. This is different from conventional face recognition which usually assumes that multiple samples are available for each subject to be recognized.

Face recognition in general has advanced significantly in the past decades [186]. Numerous commercial face recognition systems have been released and a great number of papers have been published in journals and conferences dedicated to the related area of face recognition. Early works carried out on face recognition are mostly based on the locations of fiducial facial landmarks. Geometric relationships among the facial landmarks such as the width of the head and the distances between eye corners are used to represent a face image, standard pattern recognition techniques are then employed to match faces using these geometric measurements [13, 59, 58, 94, 24]. Due to the difficulties of precisely locating all the facial landmarks and deficiency of the geometric features, later face recognition methods focus on the appearance (pixel intensities) of the face image. The appearance based methods are typically with a common goal to obtain a low-dimensional feature space for face representation, in such a way that the intrinsic characteristics of the original face samples are well preserved. The feature space can be either manually defined or learned from training images. In the first case, a variety of image transforms such as Gabor wavelet transform [158], discrete cosine transform (DCT) [42], discrete Fourier transform (DFT) [70], multiresolution wavelet transform [20] and Haar transform [113] are applied on the whole face image or local facial regions for feature extraction. Recently, with local binary pattern (LBP) being successfully applied in face recognition [3], a number of local descriptors are defined and used to extract features on face images for recognition, including variants of LBP [143, 184, 159], scale-invariant feature transform (SIFT) descriptor [98], histogram of oriented gradients (HOG) [4], histogram of Gabor phase patterns (HGPP) [133], local Gabor XOR pattern [164], Weber local descriptor (WLD) [16], etc.

Subspace analysis techniques are often employed for learning the feature space from
training samples. The subspace analysis techniques learn bases of a subspace of the image space through training, so that a face image can be represented as a combination of the bases (basis images). Dimension reduction is achieved by limiting the number of bases and characterization of the facial pattern is achieved by imposing constraints on the property of bases. The first and most representative subspace face recognition method is the principal component analysis (PCA, also known as eigenspace) [149], which imposes orthogonality constraint on the subspace learning and only selecting principal bases that account for majority of the variability in the training data. Another representative work is the Fisher linear discriminant analysis (LDA, also known as FLDA, FDA, Fisherfaces) [7]. Unlike the unsupervised learning approach adopted in PCA, LDA utilizes the label information during training and learn non-orthogonal bases of a subspace. The characteristic of the learned subspace is that training samples from different subjects are far from each other while samples from the same subjects are close to each other. Following PCA and LDA, with different criteria for subspace learning, other subspace methods have been subsequently proposed and applied to face recognition. These include unsupervised methods such as independent component analysis (ICA) [102], locally salient ICA (LS-ICA) [62], non-negative matrix factorization (NMF) [74] and locality preserving projections (LPP or laplacianfaces) [46]. Some of the supervised methods include fractional-step LDA (F-LDA) [87], direct LDA (D-LDA) [178], weighted LDA [86] and LDA with extended optimization criterion [173]. To deal with the nonlinearity in face recognition, the kernel trick [103] is widely employed to extend the linear subspace methods to nonlinear variants such as kernel PCA [83], kernel direct discriminant analysis (KDDA) [89], kernel Fisher discriminant analysis (KFD) [167] and kernel Fisher analysis (KFA) [84]. Efforts are also made to directly learn the nonlinear subspaces, these methods are often referred to as nonlinear manifold learning methods. Well-known manifold learning methods are isometric feature mapping (ISOMAP) [145], local linear embedding (LLE) [127], Laplacian Eigenmap [8], unsupervised discriminant projection (UDP) [168], and orthogonal neighborhood preserving projections [65]. Recently, methods have been proposed to achieve sparse representations [162] by learning a subspace with over-complete bases and imposing sparseness constraints (often cast into an $l^1$ minimization problem) during
2.2. Single sample-based face recognition

the subspace learning and projection [161, 182, 171, 154]. This kind of sparse representation has shown significant potential for face recognition.

Reliable recognition can be achieved by most of the above mentioned methods in controlled environments. However, face recognition under realistic and unconstrained conditions remains challenging and is far from being solved. This is mainly due to the fact that facial appearance is easily affected by the variations of pose, illumination, expression, occlusion and several other factors. One common strategy adopted by many existing face recognition methods to deal with the variations is to use multiple representative training samples for each subject. By “multiple”, we imply that the number of samples per subject should be proportional to the dimensionality of the features used for the recognition, and by “representative”, we imply that every test face image should ideally have its corresponding training sample taken under the same or similar condition. Unfortunately, the possible combinations of the mentioned variations would require an exponential number of samples to be captured for each subject. Such a requirement is hardly met in most applications. In contrast, there is often only one sample image available for each subject of interest in many applications; for instance, searching for a person of interest in video captured by surveillance cameras.

Recognition of faces based on one single sample per subject is even more challenging because intra-person variations cannot be estimated and even inter-person variations may not be reliably estimated either. This makes most conventional face recognition methods that rely on multiple training samples per person, such as Linear Discriminant Analysis (LDA) [23], and Bayesian matching methods [101], to be inapplicable or fail to perform satisfactorily. Nevertheless, the problem has recently been actively studied using different approaches and a number of techniques have been reported (see [142] for a survey).

In this section, the existing methods for single sample-based face recognition are examined with regard to how they use and exploit the information from the limited available samples. Specifically, we categorize the methods into following approaches: face to face matching, use of virtual samples, use of the single sample of other subjects and use of generic face samples.
2.2. Single sample-based face recognition

2.2.1 Face to face matching

Face to face matching is a straightforward approach to address the problem of recognizing faces from single samples. It considers the samples of different subjects independently and samples of the other subjects do not carry any information for one specific subject. Given a face image to be recognized, methods that fall into this approach in general extract the information (features) from the face image and compare the face image with the sample images (one per subject) in the database or gallery sequentially. The face image is recognized as the subject whose sample image in the gallery gives the minimum or maximum matching score (depending on how the score is defined) or rejected as an unknown subject if the minimum or maximum matching score is over or below a threshold.

The key to the success of these methods is to find and extract discriminative features for the comparison. Features extracted from local facial regions appears to be most promising. The local regions could simply be regular partitions of the image or be selected around key facial points. Early methods use geometrical features such as the width of the head and the distances between the eyes. Although these facial features were often manually extracted, the reported performance of geometric-based features is fairly rudimentary. One of the reasons is that geometrical features alone are not sufficient to fully represent a face. Textural information also contribute to the identification of a person as revealed by the cognitive study.

Manjunath et al. [93] proposed to detect key facial feature points and store two kinds information for each detected feature point. One is the location of the feature point and the other is the spatial and angular distances from the point to its neighbours. To model the relationship among the feature points, a topological graph is constructed and face matching is then formulated as a graph matching problem. The problem of this method is that once the topology graph is constructed no more modification is allowed. Such a fixed topology graph is often too rigid to accommodate the possible variations of the test face images due to varying head pose, facial expression and illumination.

The Elastic Bunch Graph Matching (EBGM) method [158] solved the above problem by proposing a deformable topology graph matching scheme. In this scheme, each face
is represented as an EBGM template graph, of which each node is attached with one or several Gabor jets that consist of the responses of different Gabor wavelets extracted at a predefined facial feature point. Scale and position of each node is allowed to vary according to the appearance variations on a specific face. However, one obvious disadvantage of the EBGM method is that only information at predefined landmark points of the face is used for recognition. Although this is a crucial factor that contributes to the robustness of the method, it cannot effectively handle the situations when the key points are occluded.

Kepenekci et al. [61] proposed a method to deal with the occlusion. In the method, instead of predefined number of feature points in a given face image, they used a set of Gabor filters to scan local facial regions, and pixels with high responses of Gabor filters are automatically chosen to be the feature points. Since the resulting feature points can be different from face to face, the possibility of finding subject-specific features is increased. For each feature point, besides the Gabor response values, its location is also recorded, thus the spatial structure of the face is implicitly encoded. In the matching process, the similarity scores of each pair of corresponding local features are summed up, the label of the training image with the largest total similarity score is then assigned to the given test image. The problem with this method is that the selected facial points with high responses of Gabor filters are not necessarily important for the recognition task. Furthermore, due to the way feature points are detected, a feature point in a test image may not have a feature point in the corresponding local area of a sample image for matching.

In [32], a new face feature representation, Line Edge Map (LEM), was proposed to integrate the structural information with spatial information of a face image by grouping edge pixels to line segments. After extraction of the single-pixel-edge map of a face image, a polygonal line is fitted to the edge map to generate the LEM of the face (see an illustration of a face LEM in Figure 2.2). Then a disparity measure, Line Segment Hausdorff Distance (LHD), is employed to measure the similarity of face LEMs. The LHD is a shape comparison measure based on LEMs. It is a distance defined between two sets of lines. Unlike most shape comparison methods that require a one-to-one correspondence between a model and a test image, LHD can be calculated without explicit line correspondence to deal with broken
lines caused by segmentation error. The experimental results on AR database showed that LEM achieved a relatively high recognition rate under pose, illumination and size variations. However, its performance was low under severe distortions such as screaming and occlusion.

Similar line based method was proposed by Park et al. [115]. In their method, a face is encoded by an ordered triple known as Face-ARG model. A Face-ARG model consists of a set of nodes (line features), a set of binary relation vectors between the nodes, and a set of relation vector spaces of the nodes. A relation vector space is defined as the set of relation vectors between a node and the rest of the nodes in the node set. In order to match two Face-ARGs, the partial attributed relational graph (ARG) matching algorithm [116] was employed. In particular, the correspondence graph (as shown in Figure 2.3, the corresponding features are connected by solid lines.) of the reference Face-ARG of a training sample and the test Face-ARG is first constructed. Then, the stochastic distance between the corresponding relation vector spaces is used as the similarity measure between the two Face-ARGs. Compared with the LEM representation, a Face-ARG represents not only the local structural information but also the whole structure of a face by using the relation vector spaces. Better results than the LEM method was reported on the AR database with single neutral frontal-view face images as the training samples. However, as the algorithm is based on the invariant binary relations between features, it may fail to work or perform poorly when there is pose variation.

Ahonen et al. [2] described spatially enhanced local appearance features. In the extraction of the feature, three different levels of locality are explored, i.e., pixel-level, regional level and holistic level. The first two levels of locality are realized by dividing a face image into small regions, from which the Local Binary Pattern (LBP) histograms are extracted to capture the texture information. The holistic level of locality, i.e. the global description of the face, is achieved by concatenating the extracted regional LBP histograms. The recognition was performed using a nearest neighbour classifier with $\chi^2$ distance as a dissimilarity measure. Although this LBP-based method was not intentionally proposed for single sample-based face recognition, due to the local nature of the features and the simple nonparametric classifier, the method demonstrated good performance when one training sample was used
2.2. Single sample-based face recognition

Figure 2.2: An illustration of a face LEM [32]

Figure 2.3: An example of the Face-ARG representation and partial matching [115]
2.2. Single sample-based face recognition

for each subject.

Methods based on the features extracted from local facial areas have the potential to be less sensitive to the variations caused by viewpoints, illumination, expression and occlusion. However, one of the key challenges for such local-based methods is how to select the facial regions for feature extraction. The selection of the regions should not be too rigid in order for the methods to be robust against various variations and, at same time, the regions should be representative and the selection should be consistent and reliable. For the methods based on the facial landmarks, their performance is often sensitive to the detection of the landmarks. Reliable detection of facial landmarks itself is a hard problem. For methods that simply partition the face image into sub-regions [2, 141], their performance would be hindered by the regions that are not salient to the recognition of the face. One strategy to reduce the hindrance is to weight the regions that would contribute more to the identity of the person. Distinctiveness of the regions potentially offers an objective way to weighting the regions. In addition, incorporation of the configural information on the facial parts in the process of selecting regions could also potentially improve the performance.

2.2.2 Use of virtual samples

Recognition of faces with multiple training samples per subject has been actively studied in the past and many methods have been proposed [186, 79, 25]. The key information missing when only one training sample is available for each subject is the intra-person variations. Such information is usually required by most multiple sample-based face recognition methods. One strategy to leverage the theories and methods developed for multiple sample-based face recognition in the context of single sample-based face recognition is to extend the single training sample for each of the samples by creating virtual samples or artificially generated samples. In the symmetrical PCA, Yang et al. [172] utilized the horizontal mirror symmetry of a face image with respect to the vertical middle line of the image and decomposed each image in the training set into an even symmetrical image and an odd symmetrical version to form two training sets: the even and odd symmetrical image sets. The standard PCA was
then applied on the two virtual sets respectively for feature extraction. Based on the observation that even symmetrical principal components often take up larger energy ratio and are generally less sensitive to variations than their odd counterparts. Feature selection was conducted according to the energy ratios with the goal to select more even symmetrical features while still keeping a number of the odd ones, and the selected PCA features were used for recognition.

Another method of constructing virtual training samples from the existing samples is to use image perturbation as introduced by Martinez [96]. The method was proposed to address the issue caused by small translations between the training samples and the test samples. Specifically, a set of possible localization errors or translation vectors were first estimated. Then, by translating a given face image within the estimated translation vectors, new samples were generated. The standard PCA was applied to the multiple new samples generated for each face for feature extraction and recognition.

In [132], Shan et al. employed geometric transforms including translation, rotation and scaling, together with gray-level transforms such as simulated directional lighting and man-made noise to derive 17 novel images from each single face sample and applied Fisherfaces for recognition. Chen et al. [18] proposed to partition each face image into a set of sub-images with the same dimensionality, and considered each sub-image as a training sample. Thus multiple training samples (composed of all the partitioned sub-images) for each person were produced and the traditional FLD-based methods were applied. Similarly, Huang et al. [55] divided the face image into 5 facial blocks, and shift each of the face block one pixel up, down, left and right to generate four more extra blocks. LDA was employed for the recognition by considering all the original and generated face blocks as training samples.

An alternative method to extend the training sample set is to construct new instances from each single sample. In [69], several linear and non-linear filters were applied on each image to produce 150 instances in total. The linear filters included Prewitt, Sobel, Laplacian, Gaussian, box filter, Gabor at 5 scales and 4 orientations, and oriented filter pairs at 3 scales and 6 orientations. The non-linear filters were morphological operators including erode, dilate, anisotropic diffusion and phase congruency. Then, an oriented component analysis
(OCA) classifier was built for each type of instances and all the OCA classifiers were finally combined with a weighted linear sum method to give the final decision.

Generally, methods that use virtual images to extend the training samples are often heuristic. The artificially generated facial images are often highly correlated and are usually not able to capture the true intra-person variations that exist in real image sets. In other words, the virtual samples are essentially different from true samples. Therefore, the recognition performance of these methods can drop substantially when the test image does not show sufficient resemblance to the virtual samples.

2.2.3 Use of the training samples of other subjects

Methods that match face to face or use virtual samples are based on a common principle that samples belonging to different subjects are independent of each other. These methods do not utilize as much the inter-subject variations as should be and the comparison can theoretically be in different feature spaces for different subjects. In the approach that uses the training samples of other subjects, the inter-subject variations are emphasized by establishing a common feature space in which the test image is compared with the samples of every subject in the gallery. A typical method to construct such a common feature space is to apply the standard PCA to the samples of all subjects to learn a low dimensional subspace with maximum data variance. However, study has shown that PCA does not perform robustly in this case. Motivated by the projection method for the face detection, Wu and Zhou [163] presented a method named \((\text{PC})^3A\) to enrich the information of face space. Let \(I(x, y)\) be an intensity image of size \(m \times n\), where \(x \in [1, m], y \in [1, n], I(x, y) \in [0, 1]\), the horizontal and vertical projections of the image are defined as \(H_I(y) = \sum_{x=1}^{m} I(x, y)\) and \(V_I(x) = \sum_{y=1}^{n} I(x, y)\), respectively. It is likely that the two projections reflect the distribution of the salient facial features that are useful for face recognition. In this approach, the projections obtained were used to synthesize a projection map, defined as \(M_I(x, y) = \frac{V_I(x)H_I(y)}{\bar{J}}\), where \(\bar{J}\) is the average intensity of the image. This projection map is then combined with the original face image to form a projection-combined image (an example of the projection map and projection-combined image is shown in Figure 2.4). Finally, PCA was applied on the projection-combined version.
2.2. Single sample-based face recognition

![Figure 2.4](image1)

Figure 2.4: An example of the projection map and projection combined image in [163]

![Figure 2.5](image2)

Figure 2.5: An example of the first-order and second-order projection map and projection combined image in [17]

of the images instead of on the original face images for recognition.

Following the $(PC)^2A$ framework, Chen [17] proposed an enhanced $(PC)^2A$ method by using second-order projection-combined images as the input for PCA. Let $J(x, y) = I(x, y)^2$, the second-order horizontal and vertical projections of the image are defined as $H_{12}(y) = \frac{1}{m} \sum_{x=1}^{m} J(x, y)$ and $V_{12}(x) = \frac{1}{n} \sum_{y=1}^{n} J(x, y)$, respectively. Then, the second-order projection map is defined as $M_{12}(x, y) = \frac{V_{12}(x)H_{12}(y)}{\bar{J}}$, where $\bar{J}$ is the mean value of $J(x, y)$. An example of the second-order projection map and second-order projection-combined image are shown in Figure 2.5. Similar to $(PC)^2A$ and enhanced $(PC)^2A$, Zhang et al. [181] also proposed a method to extend PCA at the pre-processing stage. Rather than using projections maps to generate the combined images, they employed the Singular Vector Decomposition (SVD) method to derive a perturbed image from the given sample face. The derived image was then combined with the original image for the PCA-based recognition at the later stage.
In another contribution, Yang et al. [169] focused on the covariance matrix estimation under one training sample rather than adopting a pre-processing strategy. They proposed a 2DPCA method which used 2D images to estimate the covariance matrix instead of serializing a 2D image into a 1D vector. Specifically, let each face image \( I_j (j = 1 \ldots N) \) be a \( m \times n \) random matrix (\( N \) is the total number of training samples), and the average image of all training sample be \( \bar{I} \), then the image covariance matrix \( G_t \) can be evaluated according to

\[
G_t = \frac{1}{N} \sum_{j=1}^{N} (I_j - \bar{I})^T (I_j - \bar{I}).
\]

By maximizing the criterion function: \( J(X) = X^T G_t X \), a set of optimal projection axes \( \{X_1, X_2, \ldots, X_d\} \) can be obtained. These projection axes then formed a feature space for recognition.

Another representative work that reported on how to learn a common space from the single training samples of all subjects is the paper by Tan et al. [141] wherein the self-organizing maps (SOM) is adopted. Specifically, each training face image \( I \) is partitioned into \( M \) different local sub-blocks \( R_i || M_i = 1 \ldots M \) first, then a SOM network is trained using all the sub-blocks from all the available training images irrespective of the subjects. After the SOM map has been trained, each sub-block \( R_i \) of the same face image \( I \) is mapped to its corresponding Best Matching Units (BMUs) by a nearest neighbor strategy, whose location in the 2D SOM topological space is denoted as a location vector \( l_i = \{x_i, y_i\} \). All the location vectors from the same face can be grouped as a set, i.e., \( I = \{l_i\}_{i=1}^{M} = \{x_i, y_i\}_{i=1}^{M} \), which is called the face’s “SOM-face”. Tan et al. claimed that even with single sample which is unable to faithfully represent the underlying distribution, the SOM-based representation can still extract the significant information of local facial features due to its use of the training samples of other subjects in the construction of the SOM map.

The underlying assumption of the methods using the training samples of other subjects is that the intra-subject variations of all subjects are similar and probably share a similar distribution to the distribution of all training samples. Such assumption is acceptable when there is little variations among the training samples and possible test images. However, the assumption will be violated if the face images, training or testing, have high degree of variability in pose, illumination, expression or occlusion, and thus lead to a poor performance.
2.2. Single sample-based face recognition

2.2.4 Use of generic samples

The extensive study of multiple sample-based face recognition and the decreasing cost of capturing face images have led to many free publicly available face datasets. These datasets provide rich face samples of many races, at various ages and under different conditions. Their availability has driven researchers to begin to look into strategies of how to use what is referred to as generic face samples to improve the performance of existing face recognition systems, especially single sample-based variety.

Generic face samples refer to the samples of the subjects that are not intended to be recognized. One approach to using the generic dataset is through knowledge transfer or transfer learning [112, 100, 77]. The basic idea of transfer learning is to learn knowledge from one task domain and use it in a different task domain. It includes the forms of transferring knowledge of instances, transferring knowledge of features representations and transferring knowledge of model parameters. For example, Wang et al. [56] presented a learning framework for single sample face recognition. In the framework, a feature subspace is first extracted from a generic dataset. Then all faces are projected into the generic feature subspace and face recognition is conducted based on the nearest neighbour classifier. Using the framework as a platform, they examined the performance of the feature subspaces extracted using PCA, kernel-PCA, FLDA, and regularized discriminant analysis (RLDA) and verified the positive contributions of the generic dataset. Similarly, Kim and Kittler [63] proposed a solution to the pose-invariant face recognition problem from a single frontal face image by collecting a generic training set to extract a pose-invariant subspace. In [139], Su et al. proposed an adaptive generic learning (AGL) method to estimate the intra-personal variations and the class mean for each person (with only a single sample) in the gallery. The estimation is obtained based on generic samples through least square regression. With the predicted intra-personal variations and mean of each enrolled person, the overall within-class and between-class scatter matrix can be obtained for the persons in the gallery. This thus makes FLDA applicable to the single sample face recognition. The transfer learning based methods are all use the assumption that different persons share certain common properties, thus related information
or models learned from other individuals are also valid for a new person. However, this assumption does not always hold, especially for persons of different skin colors, races or ages. As a result, negative transfer may happen, which may render the recognition performance worse than that without using knowledge transfer.

Methods have also been proposed to represent a given face image by the similarities between the face and each of labeled generic face samples at hand. By treating similarities as features, direct access to the features of each face image is not required, and thus the feature space for similarity comparison could be arbitrary. However, this kind of representation may not capture discriminative information if there is a large intra-personal variations compared to the inter-personal variations in the generic face set. These similarity-based methods can be extended by training a number of well designed specific classifiers based on the labeled generic samples and representing a given face image by the outputs of the pre-learned classifiers. Based on this “classifier stacking” based representation, Bart et al. proposed a single-example learning framework, and applied the framework to face recognition. Kumar et al. proposed to use pre-learned classifiers that correspond to both high level and low level facial features. The high level feature classifiers were referred to as attribute classifiers, each of them was trained to recognize the presence or absence of a describable aspect of visual appearance of face such as “male”, “white”, “mustache”, “curly hair”, “smiling” and “chubby”. The low level feature classifiers were referred to as similar classifiers, each of them was trained to recognize the similarity between one facial component (such as eyes, nose, and mouth) and a specific kind. The problem with the “classifier stacking” based representation is that the cost of collecting and labeling the training samples for pre-learned classifiers may be prohibitively high, and the efficiency of this representation is highly related to the performance of each pre-learned classifier.

A recent work of Wolf et al. utilized generic face samples (referred to as background samples in [159]) in a different manner. In their work, by considering the generic samples as negative training examples, discriminative models are trained to measure the similarity of two given face images; one is usually the test face image and the other is a sample face from the gallery. In particular, the method, called one-shot similarity, learns one model
for each of the given face images using the face image itself as a single positive training sample and the generic samples as the negative ones. It then measures the similarity of the two faces based on the average of the two confidence scores obtained by using one model to recognize the other face. In the method called two-shot similarity, a single model is trained using both given faces as positive training samples and the generic samples as the negatives ones, and the similarity is measured based on how well the learned model discriminates the positive and negative sets. In general, the nature of both the one-shot and two-shot similarity depend on how the models are trained based on limited positive training samples and what kind of classifier is used. Although good results are reported in [159] by employing LDA and SVM classifiers, the generalization for one-shot and two-shot similarity is still an open issue.

2.3 Age estimation

Age estimation through the analysis of the visual appearance of face images has many potential applications including demographics, customer group analysis and image/video content management. However, it has been much less studied compared to face recognition and facial expression recognition. One of the key reasons is that the personalized and uncontrolled aging process makes age estimation a challenging task even for human beings. Another reason is that large datasets for the study of age estimation are hard to collect. With the recent release of the databases, MORPH [125], FG-NET[1], that were specifically designed for age estimation, the problem has been actively studied from two perspectives: age group estimation, in which estimating the range of the ages that a subject may fall into is the goal; and exact age estimation, with the goal of estimating the exact age of a subject. Exact age estimation is often treated as a regression problem, while the age group estimation can be formulated as a classification problem. In this section, we categorize the existing age estimation methods into two broad approaches namely, geometric-based and appearance-based, and provide a brief review of the state-of-the-art methods.
2.3. Age estimation

2.3.1 Geometric-based approach

The geometric-based approach is basically based on the observation that while a subject gets old, the ratios of the distances between his/her facial landmarks such as eye corners, mouth corners and nose tips changes. Methods that belongs to this approach detect from a face image these facial landmarks either manually or automatically and measure the ratios of the distances as features to determine the age of the subject. Kwon and Lobo [68] proposed an age classification method based on the studies in craniofacial development theory that uses a mathematical model to describe the growth of a person’s head from infancy to adulthood. They computed six ratios of distances on frontal face images and used them to separate babies from adults. Ramanathan and Chellappa [122] later used eight ratios of distance measures to model age progression in young faces from 0 to 18 years.

Obviously, the geometric-based methods are sensitive to the localization of the facial points and mainly work for frontal face images. Furthermore, the shape of human faces do not change too much during adult ages, thus without considering the texture information, geometric-based methods may work for young ages but are not suitable for adults.

2.3.2 Appearance-based approach

The appearance-based approach focuses on extracting edge, shape and texture features that are related to aging. Hayashi et al. [45] proposed to use both texture and shape features together with a semantic-level descriptor to represent each face image. Their system estimates human ages through a multiple-group classification scheme, each group covering five years. Gender estimation was also employed to improve the age estimation considering the fact that aging patterns are different between males and females.

The active appearance models (AAMs) [140] is a statistical face model proposed initially for coding face images. Given a set of training face images, a statistical shape model and an intensity model are learned separately based on the principal component analysis. By modelling the face with both shape and texture information, AAMs has been shown to be efficient in many applications and was also used as the face representation for age estimation.
2.4. Head pose estimation

Based on AAMs, Lanitis et al. [72] proposed an aging function, \( \text{age} = f(b) \), where \( \text{age} \) is the actual age of an individual in a face image, \( b \) is a vector containing 50 raw model parameters learned from the AAMs. Lanitis et al. also compared different classifiers for age estimation based on AAM features in their later work [71] and demonstrated that the simple nearest neighbor classifier achieved the best performance.

With AAMs-based face encoding (200 AAMs features to encode each face), Geng et al. [34, 35] addressed the age estimation problem by introducing an aging pattern subspace (AGES), which is a subspace representation of a sequence of individual aging face images. The proper aging pattern for a previously unseen face image is determined by the projection in the subspace that can reconstruct the face image with minimum reconstruction error, while the position of the face image in that aging pattern will then indicate its age.

The problem of AAMs-based representation for age estimation is that AAMs only encode the image intensities without involving any spatial neighborhood to calculate texture patterns. Intensities of single pixels usually cannot characterize local texture information [31]. However, many studies have suggested that age features are usually encoded by local information, such as wrinkles and skin texture on the forehead or at the eye corners.

In order to extract the local information, local descriptors such as LBP [40] and spatially flexible patch (SFP) that integrates coordinate information together with the local DCT feature [166] were employed in age estimation. Because local features are generally less sensitive to face distortions, age estimation methods based on local features are able to perform better than the holistic methods, especially when a certain degree of misalignment, occlusion, and head pose variations occurred.

2.4 Head pose estimation

Head pose estimation is a very useful front-end processing for handling pose variations in automatic face analysis such as face recognition, face detection and expression recognition. The orientation of a user’s head relative to the camera is also an important factor for applications like passive navigation and human-computer interaction (HCI). Existing representative
methods for head pose estimation are briefly reviewed in this section by categorizing them into geometric-based and appearance-based approaches.

2.4.1 Geometric-based approach

In the geometric-based approach, location of key facial points and their relative configuration are often used to determine head pose. For instance, Gee et al. [33] used five facial points, the outside corners of each eye, the outside corners of the mouth, and the tip of the nose, to estimate the head pose. In their method, the facial symmetry axis is found by connecting a line between the midpoint of the eyes and midpoint of the mouth. Assuming a fixed ratio between these facial points and a fixed length of the nose, the head pose can be determined under weak-perspective geometry from the 3D angle of the nose.

Horprasert et al. [49] estimated head poses using a different set of five points (the inner and outer corners of each eye, and the tip of the nose). Under the assumption that all four eye points are assumed to be coplanar, the three degrees of freedom for a head can be estimated as follows. Yaw can be determined from the observable difference in size between the left and right eye due to projective distortion from the known camera parameters. Roll can be found simply from the angle of the eye-line from the horizon. Pitch is determined by comparing the distance between the nose tip and the eye-line to an anthropometric model.

Rather than using manually labeled facial points as in the two methods described above, the method recently proposed by Wang et al. [157] utilized the automatically detected inner and outer corners of each eye and the corners of the mouth. The method is based on the observation that three lines between the outer eye corners, the inner eye corners, and the mouth should be parallel. Any observed deviation from being parallel in the image plane is a result of perspective distortion. The vanishing point can then be calculated using a least squared error method to minimize the overdetermined solution for the three lines. This point can be used to estimate the 3D orientation of the parallel lines if the ratio of their lengths is known and it can be used to estimate the absolute 3D position of each feature point if the actual line lengths are known.

The geometric-based methods are fast and simple. However, one drawback is that they
are very sensitive to the localization of the required facial points. Furthermore, these methods are all based on the assumption that the configuration of facial points does not change significantly under different facial expressions, which is often violated.

### 2.4.2 Appearance-based approach

Appearance-based methods avoid the problem of facial points localization, they typically use holistic face appearance as input. Some of these methods employ regression tools such as neural networks to develop a functional mapping from the face image to a head pose measurement. For example, two neural networks were trained in [185] to approximate functions that map a face image to the pan and tilt angles of the head respectively. Each of them consists of a multi-layered perceptron (MLP) with a single hidden layer. The input layer of the neural network consists of a $48 \times 48$ unit as “retina” which receives the face image. The hidden layers of the pan and tilt neural networks consist of five and six units respectively. The output layer consists of 19 units representing equally spaced angle values ranging from $0^\circ$ to $180^\circ$. Each of the 2304 units in the input layer is fully connected to the hidden layer, which in turn is fully connected to the output layer. To determine the head pose, the image is resized to a $48 \times 48$, and projected onto the input layer of each neural network. After completing a forward pass through each network, the pan and tilt angles are derived from the output activation levels using a least squared error fitting.

Rae et al. [120] presented a head pose estimation method based on locally linear map (LLM) neural networks. The basic idea underlying the LLM network is to represent the to-be-learnt non-linear mapping as a collection of locally valid linear mappings that are learned by separate units. The LLM-network used in [120] is a feedforward network consisting of a single internal layer with nodes. Every node receives the same input vector (each face image is represented as a vector of Gabor features). The output of each single node is obtained through a local linear mapping that is defined by an input weight vector, an output weight vector and a Jacobian matrix. The matrix denotes an approximation of the local gradient of the mapping function around the input weight vector. Given the input vector, a node is selected by minimizing the distance between input vector and the input weight vectors;
then the output of the selected node is considered as the estimated head pose. According to
the survey conducted by Murphy-Chutorian et al. [104], the regression-based methods have
performed favourably in head pose estimation. However, these methods often require a large
number of training data covering all the head poses with accurate labels. In practice, it is
difficult to collect such large and accurate training sets.

Other appearance-based methods formulate the head pose estimation as a pattern classi-
fication problem. The range of head orientations is divided into a limited number of classes
and classifiers are trained. Rather than directly using the whole face image as the input
feature vector to the classifiers, these methods usually employ subspace analysis (or mani-
fold learning) techniques such as principal component analysis (PCA), linear discriminant
analysis (LDA), independent component analysis (ICA) [80] and isometric feature mapping
(Isomap) [123] to extract a low-dimensional feature vector from each face image for classi-
fication. However, the features in the subspace are often corrupted by the variations caused
by irrelevant factors such as identity, expression and illumination. Thus, the key challenge
of appearance-based methods lies in constructing a feature subspace that is sufficiently dis-
criminative with respect to the head pose while ignoring other sources of image variations.


c 2.5 Databases for face analysis

Along with the study of face analysis, many datasets have been established and made publicly
available for evaluating and benchmarking new algorithms. Since face analysis is heavily
impacted by many factors such as illumination, occlusion and expression variation, each
dataset was often created for a specific face analysis task and under a controlled environment.
In this section, we list the commonly used datasets for face recognition, age estimation and
head pose estimation, and summarize the conditions under which these datasets were created
and their originally intended applications.
2.5. Databases for face analysis

Figure 2.6: Sample images from the AR face database

Figure 2.7: Sample images from ORL Face database

2.5.1 Databases for face recognition

- **The AR Face Database** [95]: The database contains over 4,000 colour images corresponding to the faces of 126 people (70 men and 56 women). The pictures were taken at the Computer Vision Centre of University of Alabama at Birmingham under strictly controlled conditions. Each of the subjects in the database were recorded twice at a two-week interval. During each session 13 conditions with varying facial expressions, illumination and occlusion were captured. No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed on participants. Sample images are shown in Figure 2.6.

- **The ORL Database** [129]: The database contains face images of 40 distinct subjects. For each subject, ten different images are provided. The images were taken at different times, varying (not systematically) the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for side movement). Sample images are shown in Figure 2.7.
2.5. Databases for face analysis

- **The BioID Face Database** [30]: The dataset consists of 1521 gray level images with a resolution of 384x286 pixels. Each one shows the frontal face of one of the 23 subjects. All the images are taken in uncontrolled conditions using a web camera in an office environment. For each image in the database, the ground truth of the 20 facial feature points were obtained through manual annotation and are supplied with the database. Sample images are shown in Figure 2.8 and the facial feature points are presented in Figure 2.9.

- **The LFW database** [54]: The LFW database contains about 13,233 images of faces collected from the web. Each face has been labeled with the name of the person pictured. The dataset has two versions: the original version and the funneled version, in which images are automatically aligned. Sample images are shown in Figure 2.10.

- **The FERET Database** [119]: The database contains 24 subsets of gray-scale facial images which were recorded in 15 sessions between August 1993 and July 1996. Five frontal image subsets of the database are widely used for face recognition evaluation. Sample images of the five subsets are shown in Figure 2.11. Among the five subsets, images of subset $fa$ and subset $fb$ were obtained in close succession. The subjects were asked to display a different facial expression for subset $fb$. The images in subset $fc$ were recorded with a different camera and under different lighting. A number of
2.5. Databases for face analysis

Figure 2.10: Sample images from LFW Face database

Figure 2.11: Sample images of the five frontal image subsets of FERET database

subjects returned at a later date to be imaged again. For the images in the *duplicate I* subset, 0 to 1031 days passed between recording sessions. For the images in the *duplicate II* subset, at least 18 months separated the sessions.

### 2.5.2 Databases for age estimation

- **The FG-NET Aging Database** [1]: The FG-NET aging database is publicly available. It contains 1,002 high-resolution color or gray-scale face images of 82 multiple-race subjects with large variation of lighting, pose, and expression. The age range is from 0 to 69 years with chronological ageing images available for each subject (on average, 12 images per subject). The aging features can be either a shape model with 68 key points, the AAM with 200 model parameters, or the appearance. Sample images are
2.5. Databases for face analysis

Figure 2.12: Sample images from FG-NET Face database

Figure 2.13: Sample images from MORPH Face database

shown in Figure 2.12.

- The MORPH Database [125]: The MORPH face database was collected by the Face Ageing Group at the University of North Carolina at Wilmington for the purpose of face biometrics applications. This longitudinal database records individuals’ metadata, such as age, gender, ethnicity, height, weight, and ancestry, which is organized into two albums. Album 1 contains 1,724 face images of 515 subjects taken between 1962 and 1998. The ages range from the average of 27.3 to maximum 68 years. There are 294 images of females and 1,430 images of males. The age span is from 46 days to 29 years. Album 2 contains more than 20,000 face images obtained from more than 4,000 individuals. Sample images are shown in Figure 2.13.
2.5.3 Databases for head pose estimation

- The CMU PIE Database [134]: The CMU PIE database systematically samples a large number of pose and illumination conditions along with a variety of facial expressions. The database contains 41,368 images obtained from 68 individuals. The subjects were imaged in the CMU 3D Room using a set of 13 synchronized high-quality colour cameras and 21 flashes. The resulting RGB colour images are 640 in size. Figure 2.14 shows example images of a subject in all 13 poses. In addition to the pose sequence, each subject was recorded under different illumination and expressions.

![Sample images from CMU PIE face database](image)

Figure 2.14: Sample images from CMU PIE face database

2.6 Summary

Automatic face analysis is in general a challenging task, especially under realistic conditions. Firstly, it is because different types of information are mixed and embedded in a face. When only one type of information is required for a specific face analysis task, the other types of information become negative factors that could severely interfere with the required information. Secondly, the facial appearance can also be easily changed by environmental factors such as illumination and occlusion. Extraction of the required information from the mixture seems very difficult as evidenced by the extensive study in the past. In contrast, humans are able to extract the information in a selective and effortless manner and this may be partially explained by some of the recent cognitive findings related to face perception as summarized in Section 2.1. This thesis aims to develop computational models that account for the findings. It is expected that such computational models will provide a set of new tools
for face analysis.
Chapter 3

Localized compact face representation

In the past, a number of features have been defined intuitively to represent face images in the ways that simulate different aspects of human visual perception. For example, motivated by the distribution of the receptive fields in the human retina, Bicego et al. [12] sample each face image using patches derived from a log-polar mapping. In the work of Gu et al. [38], local multi-scale Gabor-filter operations are first applied to face images and the resulting Gabor decompositions are then encoded using radial grids to imitate the topographical map-structure of the human visual cortex. Different from these biologically driven features that are manually defined, the abstract reusable local features suggested by Wallis et al. are learned from example faces. One widely used tool for such learning is non-negative matrix factorization (NMF) [74]. However, NMF dose not necessarily produce effective localized representations, especially when training samples are not properly aligned. Therefore, in this chapter, NMF is extended for robust learning of localized compact representation of face images. Such a localized compact face representation is not only in line with the reusable local features suggested by Wallis et al., but also robust against variations such as partial occlusions and local distortions. The extended NMF (ENMF) is solved by a projected gradient algorithm with a data-driven initialization scheme. In addition, an indicator is proposed to objectively measure the locality and compactness ($LC$) of a localized representation and to quantitatively evaluate the efficiency of ENMF. Experimental results on benchmark face databases shown that the proposed ENMF is much more effective in learning localized compact representation and more tolerant to the variations, especially, misalignment of the training samples, than conventional NMF and its major variations.
3.1 Non-negative matrix factorization

Non-negative matrix factorization (NMF) is a linear, non-negative approximate data representation. Given a non-negative data matrix \( V = (v_{ij})_{m \times n} \), NMF finds non-negative matrices \( W = (w_{ij})_{m \times r} \) and \( H = (h_{ij})_{r \times n} \), such that \( V \approx WH \). The rank \( r \) of the factorization is generally chosen to satisfy \((n + m)r < mn\), so that the product \( WH \) can be regarded as a compressed form of the data in \( V \). Let \( V \) be a face database, each column of \( V \) contains \( m \) pixel values of one of the \( n \) face images in the database. Then, each face in \( V \) can be represented by a linear combination of \( r \) columns of \( W \), known as basis vectors (images). Each column of \( H \) is called a coefficient vector, that is in one-to-one correspondence with a face in \( V \), and describes how strongly each basis is present in the face.

NMF is often considered as an optimization problem. \( W \) and \( H \) are chosen to minimize the reconstruction error between \( V \) and \( WH \), which is usually measured by Euclidean distance. The objective function can be written as:

\[
E(W, H) = \| V - WH \|^2 = \sum_{i,j} (V_{ij} - (WH)_{ij})^2.
\]  

(3.1)

This minimization problem is not convex with respect to both \( W \) and \( H \), but convex with respect to each variable separately. Paatero and Tapper [111] proposed a gradient descent method to solve this problem and, Lee and Seung [75] devised a multiplicative algorithm to search a local optimum.

Several variants of NMF have been proposed to improve NMF from different perspectives. For example, Zafeiriou et al. [179] proposed discriminant NMF (DNMF) and Yang et al. [170] proposed nonnegative graph embedding (NGE) to introduce discriminant information into NMF for better classification power; To improve the efficiency of NGE, multiplicative NGE (MNGE) [156] is presented by Wang et al. and projective NGE (PNGE) [85] is further developed by Liu et al; Cai et al. [15] and Zhang et al. [183] proposed graph-regularized NMF (GNMF) and topology preserving NMF (TPNMF) respectively to account for geometric structure in the data; Guan et al. introduced regularized discriminative (NMF MD-NMF) [39] that takes both the geometry of data and the discriminative information of
different classes into consideration. In this thesis, we focus on the locality of NMF basis images. Since only additive combination of the bases are allowed, NMF is considered to be compatible with the intuitive notion of combining parts to form a whole in an accumulative manner. The learned basis images tend to be local part-based. However, due to the non-convex nature and variations in the training samples, the solution of NMF in most cases is neither unique nor necessarily part-based for real face datasets. Donoho et al. [26] theoretically proved that NMF would produce a unique part-based decomposition only when learning from datasets that obey certain conditions, that are:

- Each image in the database can be represented as a linear combination of the different parts in the different positions. Both parts and weights of the linear combination obey the non-negativity constraint.
- The different bases are linear independent.
- The database contains all combinations of parts in the different positions.

Such conditions are quite restrictive and hardly satisfied in real life datasets, especially for facial image databases. It is not feasible to have all the possible images with combinations of different eyes, noses, mouths in different positions. One way to deal with this realistic difficulty is to apply NMF on well aligned face datasets, in which all the eyes, noses and mouths are in the same position, so that an approximation of the decomposition into parts can be obtained with the assumption that there is only one possible position for these facial parts. However, NMF is sensitive to the misalignment and often unable to learn an effective local representation. For example, as reported by Li et al. [78], when NMF is applied to the ORL face database [129] in which faces are not properly aligned, the learned basis images are holistic rather than local part-based. To address this issue, Li et al. suggested the enforcement of maximum sparsity in $H$ and maximum orthogonality in $W$ and proposed the local NMF method (LNMF)[78]. Hoyer [50] adopted a different approach by explicitly controlling the sparseness of $W$ and $H$ in a method known as, NMF with sparseness constraints (NMFsc).

Both LNMF and NMFsc improved the localization of the basis images to some extent. For LNMF, maximum sparsity in $H$ enforced the condition that a basis component cannot
be further decomposed into more components. Thus the overlap between basis images was reduced. However, a high sparseness in the columns of $H$ forces each coefficient to represent more of the image, thereby making the basis images global. Consider an extreme case when only one element in each column of $H$ is allowed to be non-zero, then the NMF reduces to vector quantization (VQ), and all the basis images become holistic prototypical faces. At the same time, high sparseness in the rows of $H$ causes each learned basis to be present in a very small fraction of the training images. Thus, the learned bases tend to be redundant; a conflict with the purpose of the orthogonality constraint. In NMFsc, a particular sparseness for the columns of $W$ requires that each learned basis image has a certain fraction of pixels with values greater than zero. However, for a very high sparseness, the small fraction of non-zero pixels are not necessarily locally distributed in the basis image. Therefore, by only imposing sparseness constraint, NMFsc does not give a better part-based representation than LNMF.

3.2 Extended NMF

To overcome the shortcomings of LNMF and NMFs, the proposed extended NMF (ENMF) imposes orthogonality constraint on the basis matrix $W$ so as to reduce the overlap between basis images and explicitly controls the sparseness of coefficient matrix $H$ in order to ensure a balance between localization and overlap of the basis images. Let $U = W^T W$. We propose to impose the orthogonality constraint by minimizing $\sum_{i,j, i \neq j} U_{i,j}$. The objective function of the ENMF can be written as:

$$E(W,H) = \|V - WH\|^2 = \frac{1}{2} \sum_{i,j} (V_{ij} - (WH)_{ij})^2 + \alpha \sum_{i,j, i \neq j} U_{i,j},$$

(3.2)

where $U = W^T W$, $\alpha$ is a small positive constant. Then the ENMF is formulated into the following optimization problem:

$$\min_{W,H} E(W,H) \quad s.t. \quad W, H \geq 0, \sum_i W_{ij} = 1 \quad \forall j, \quad \text{sparseness}(h_j) = S_h \quad \forall j,$$

(3.3)
where \( h_j \) is the \( j \)-th row of \( H \); \( S_h \) is the desired sparseness of \( H \); the sparseness is measured based on the relationship between the \( L_1 \) norm and the \( L_2 \) norm [50]:

\[
\text{sparseness}(h_j) = \frac{\sqrt{n} - \|h_j\|_1 / \|h_j\|_2}{\sqrt{n} - 1},
\]

(3.4)

where \( n \) is the dimensionality of row vector \( h_j \). This measure quantifies how much energy of the vector is packed into a few components. This function evaluates to 1 if and only if \( h_j \) contains a single non-zero component. Its value is 0 if and only if all components are non-zero and equal.

A projected gradient descent algorithm [22] [9] is used to find a solution to the objective function (3.2). The algorithm updates \( W \) and \( H \) alternately. At each iteration, \( W \) and \( H \) are first updated along the direction of the negative gradient to reduce the objective function, then the updated \( W \) and \( H \) are projected onto the constraint space defined by (3.3).

In particular, \( W \) is updated by the following rules:

\[
W_{ia} \leftarrow W_{ia} - \epsilon_{ia} \nabla W E(W,H)_{ia},
\]

(3.5)

\[
W_{ia} \leftarrow P_W(W_{ia}),
\]

(3.6)

where projection \( P_W(\cdot) \) is a simple normalization to constrain the columns of \( W \) to sum to unity,

\[
P_W(W_{ia}) = \frac{W_{ia}}{\sum_j W_{ja}}.
\]

(3.7)

This is a convenient way of eliminating the degeneracy associated with the invariance of \( WH \) under the transformations and is widely adopted in representative NMF algorithms. \( \epsilon_{ia} \) in (3.5) is a step size and allowed to change at every iteration. Step size is a crucial parameter of a gradient descent algorithm. When it is too small the algorithm exhibits slow convergence and when it is too large the algorithm may diverge. A commonly used strategy is to select search for a step size that satisfies certain conditions [10, 82]. This search operation is usually time consuming and does not guarantee an optimal value. In our algorithm, we directly set the step size to
\[ \epsilon_{ia} = \frac{W_{ia}}{(WHH^T)_{ia} + \alpha \sum_i W_{ia}}. \]  
(3.8)

Then (3.5) can be rewritten as:

\[ W_{ia} \leftarrow W_{ia} - \epsilon_{ia} \left[ (WHH^T)_{ia} - (VH^T)_{ia} + \alpha \sum_i W_{ia} \right] = W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia} + \alpha \sum_i W_{ia}}. \]  
(3.9)

We can see that the update rule of \( W \) is essentially a multiplicative algorithm [76], thus the non-negativity is ensured and the convergence of such algorithm has been proved in [76]. However, as suggested by Lin [21], the proof in [76] has not shown that the multiplicative updates converge to a stationary point when the numerator of (3.8) is zero and the gradient \( \nabla W E(W, H)_{ia} < 0 \). Therefore, following the work of Lin [21], we set \( \epsilon_{ia} \) in this thesis as:

\[ \epsilon_{ia} = \frac{\overline{W}_{ia}}{(WHH^T)_{ia} + \alpha \sum_i \overline{W}_{ia}}, \]  
(3.10)

where

\[ \overline{W}_{ia} = \begin{cases} W_{ia} & \text{if } \nabla W E(W, H)_{ia} \geq 0, \\ \max(W_{ia}, \zeta) & \text{if } \nabla W E(W, H)_{ia} < 0, \end{cases} \]

\( \zeta \) is a pre-defined small positive number.

Corresponding to (3.5) and (3.6), \( H \) is updated by the following rules:

\[ H_{aj} \leftarrow H_{aj} - \eta_{aj} \nabla H E(W, H)_{aj} = H_{aj} - \eta_{aj} \left[ (W^T WH)_{aj} - (W^T V)_{aj} \right], \]  
(3.11)

\[ H_{aj} \leftarrow P_H \left( H_{aj} \right). \]  
(3.12)

Unlike the orthogonality constraint on \( W \), the sparseness constraint on \( H \) is imposed through the projection step. Thus, to further reduce the complexity of gradient descent step, the step size \( \eta_{aj} \) is initially set to 1, then multiplied by a factor \( \psi \) at each subsequent iteration. By choosing a proper small value of \( \psi \), the convergence can be guaranteed [9] (discussions about the choice of \( \psi \) are given in Section 3.3.3).

The projection \( P_H(\cdot) \) projects each \( n \)-dimensional row vector \( h_j \) to the closest non-negative row vector \( h'_j \) with unit \( L_2 \) norm, and \( L_1 \) norm being set to \( \lambda \) according to the desired sparseness \( S_h \):
\[ \lambda = \sqrt{n} - S_h(\sqrt{n} - 1). \tag{3.13} \]

Specifically, the projection \( P_H(\cdot) \) follows the following steps:

1. Set \( t \leftarrow h_j + (\|h_j\|_1 - \lambda/n)e \) with \( e = [1, \ldots, 1]^T \in \mathbb{R}^n \).

2. Set \( y \leftarrow (\lambda/n)e \).

3. Set \( h_j' \leftarrow y + \beta (t - y) \) with \( \beta > 0 \) such that \( \|h_j'\|_2 = 1 \).

   - if the \( i \)-th element of \( h_j' \) is negative then

4. Fix the \( i \)-th element of \( h_j' \) to zero.

5. Remove the \( i \)-th element of \( h_j \).

6. Decrease dimension \( n \leftarrow n - 1 \).

7. Go to 1.

Let \( W^p H^p \) denote the updated \( W \) and \( H \) after \( p \) iterations, the iterative algorithm described by (3.5) (3.6), (3.11) and (3.12) is stopped if one of the following conditions is met:

- \( \|\nabla E(W^p, H^p)\|_F \leq \tau \|\nabla E(W^1, H^1)\|_F \),

- no change in the updates for \( W \) and \( H \), or

- the number of iterations achieves the predefined maximum number of iterations.

The cost of each iteration of the proposed projected gradient descent algorithm for solving the ENMF optimization is \( O(nmr) \). By successfully avoiding the step size search process, the overall complexity of our algorithm is only related to the size of data matrix \( V \), the factorization rank \( r \) and the total number of iterations. Because the rate of convergence of the algorithm depends on the initialization of \( W \) and \( H \), we propose a new initialization method to reduce the required number of iterations. Recall that each column of \( V \) represents one image in the dataset and each row of \( V \) reflects the variations of one pixel across the samples.
Pixels can be grouped in the sample images into tentative parts based on the row vectors of $V$ and the grouped pixels can then be used to set the initial values of $W$ and $H$. This strategy will encourage the iterative process to converge quickly to an instance of local part-based representation.

Let $B = \{1, \ldots, m\}$ denote the set of indices of row vectors in $V$, $B_j$ denote a subset of $B$. Suppose all of the row vectors in $V$ can be grouped into $K$ clusters ($K$ can be directly set to $r$, the rank of ENMF), each cluster $C_i$ is associated with a submatrix $V_i (i = 1, \ldots, K)$. $V_i = V[B_i; :]$ that consists of the corresponding row vectors of $V$. Based on rank-one approximation, we have:

$$V_i \approx u_i \sigma_i p_i^T$$  \hspace{1cm} (3.14)

where $u_i$ and $p_i$ are left and right singular vectors associated with the largest singular value of $V_i$. Then the basis matrix $W$ and coefficient matrix $H$ can be initialized as:

$$W[B_i; i] = u_i$$  \hspace{1cm} (3.15)

$$W[\bar{B}_i; i] = \xi$$  \hspace{1cm} (3.16)

$$H[i; :] = \sigma_i p_i^T$$  \hspace{1cm} (3.17)

where $W[\bar{B}_i; i]$ denotes the rest of the elements in the $i$th column of $W$ excluding elements with the indices in $B_i$. $\xi$ is a small positive value replacing 0 to prevent numerical instability of the parameters under the multiplicative update rules.

### 3.3 Experimental results

Face images from ORL, BioID and FERET datasets were used to evaluate the performance of the proposed ENMF. Both visual comparison and objective evaluation were adopted in comparing the locality and compactness of the bases obtained using ENMF with those obtained using LNMF and NMFsc.
3.3. Experimental results

3.3.1 Visual comparison

For a fair comparison, NMF, LNMF, NMFsc and ENMF were applied to the original ORL database in a similar manner as reported in [78] [50]. The learned bases are shown in Figure 3.1. For NMFsc, following Hoyer’s [50] work, we set the sparseness of basis matrix, $S_w$, to 0.75 and kept $S_h$ unconstrained. For ENMF, $S_h$ was set to 0.1, $\alpha$ was set to 1, $\psi$ was set to 0.5 and $\xi$ was set to 0.001, K-Means was employed to group the row vectors in data matrix $V$ during initialization. As seen from the figure, more localized and less overlapped basis images are obtained using ENMF than NMF, LNMF and NMFsc. Facial parts like nose, eyes and mouth of different individuals can be represented by ENMF bases. Except a few bases that are associated with background, the basis images obtained by ENMF can be regarded as the “middle level facial patterns” between pixels and actual facial parts. These local facial patterns are learned from previously seen faces and “reusable”. They are in line with the abstracted features suggested by Wallis et al. [155].

3.3.2 Objective evaluation

To further quantitatively evaluate the efficiency of ENMF with respect to the capability of realizing the compact local part-based representation and robustness against misalignment of faces, an indicator, referred to as locality and compactness ($LC$) is defined. Let $r$ denote the total number of basis images (rank), $e_i$ denote the number of basis images that covers a local region $g_i$ (Figure 3.2 shows an example of ENMF bases related to a “mouth” area), assume the whole image is divided into $m$ sub-regions $g_1, \ldots, g_m$, then $LC$ is defined as:

$$LC = \frac{1}{m} \sum_{i=1}^{m} \frac{e_i}{r}.$$  \hspace{1cm} (3.18)

$LC$ is an important indicator measuring the localization and compactness of the basis images. When the basis images are holistic, the value of $LC$ is close to 1, almost all the bases contribute to a local region. A small value of $LC$ is desirable for a local compact part-based basis images. In this case, only a few bases overlap with each specific local region and local variations such as partial occlusions and local distortions only affect a small part of the
3.3. Experimental results

Figure 3.1: Basis images learned from ORL database using different methods ($r = 49$).
3.3. Experimental results

Figure 3.2: ENMF basis images learnt from normalized ORL database with $r = 81$, a group of ENMF bases related to the “mouth” area are marked with red squares.

coefficients used to represent an image. Thus, a more robust and compact representation is achieved than those cases with larger values of $LC$.

The ORL dataset and the gallery subset of the FERET dataset were used for an objective evaluation and comparison of the performance of NMF, LNMF, NMFsc and ENMF. To evaluate the robustness, the images from FERET database were transformed to simulate three possible in-plane misalignment: random translation within $[+5, -5]$ pixels in vertical or horizontal; random rotation within $[+15^\circ, -15^\circ]$; random scaling within $[0.9, 1.1]$. After the transformation, all images were cropped to $64 \times 64$ face images. For the ORL dataset, faces of the images were detected and resized to $64 \times 64$. Figure 3.3 shows some sample face images used in the experiments.

The NMF, LNMF, NMFsc and ENMF were applied to the face images with different total number of bases and same parameters as used in the visual evaluation experiment. To calculate $LC$ for each representation, the face images were equally divided into non-overlapping $8 \times 8$ sub-regions. Table 3.3.2 and Table 3.3.2 show the calculated $LC$ for the representations learned from ORL database and FERET database respectively. As can be seen from the tables, the proposed ENMF achieved the smallest $LC$ in all the cases, especially for ORL database that contains faces with out-of-plane rotations.
3.3. Experimental results

Samples of Transformed FERET face images

Samples of ORL face images

Figure 3.3: Sample face images used in the experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>$r = 36$</th>
<th>$r = 49$</th>
<th>$r = 64$</th>
<th>$r = 81$</th>
<th>$r = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>0.973</td>
<td>0.976</td>
<td>0.976</td>
<td>0.976</td>
<td>0.979</td>
</tr>
<tr>
<td>NMFsc</td>
<td>0.423</td>
<td>0.571</td>
<td>0.727</td>
<td>0.757</td>
<td>0.801</td>
</tr>
<tr>
<td>LNMF</td>
<td>0.258</td>
<td>0.238</td>
<td>0.199</td>
<td>0.195</td>
<td>0.187</td>
</tr>
<tr>
<td>ENMF</td>
<td>0.194</td>
<td>0.172</td>
<td>0.127</td>
<td>0.125</td>
<td>0.121</td>
</tr>
</tbody>
</table>

Table 3.1: LC for different representations learned from FERET database

<table>
<thead>
<tr>
<th>Method</th>
<th>$r = 36$</th>
<th>$r = 49$</th>
<th>$r = 64$</th>
<th>$r = 81$</th>
<th>$r = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>NMFsc</td>
<td>0.667</td>
<td>0.836</td>
<td>0.946</td>
<td>0.956</td>
<td>0.969</td>
</tr>
<tr>
<td>LNMF</td>
<td>0.422</td>
<td>0.396</td>
<td>0.345</td>
<td>0.341</td>
<td>0.329</td>
</tr>
<tr>
<td>ENMF</td>
<td>0.275</td>
<td>0.232</td>
<td>0.197</td>
<td>0.193</td>
<td>0.186</td>
</tr>
</tbody>
</table>

Table 3.2: LC for different representations learned from ORL database
3.3. Experimental results

<table>
<thead>
<tr>
<th>Methods</th>
<th>NMF</th>
<th>NMFsc</th>
<th>LNMF</th>
<th>ENMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU time (seconds)</td>
<td>133.1</td>
<td>350.9</td>
<td>318.1</td>
<td>237.2</td>
</tr>
<tr>
<td>Error (MSE)</td>
<td>148832.8</td>
<td>148426.8</td>
<td>4011.5</td>
<td>3774.3</td>
</tr>
<tr>
<td>Iterations</td>
<td>1267</td>
<td>1782</td>
<td>2455</td>
<td>1030</td>
</tr>
</tbody>
</table>

Table 3.3: Typical CPU times, factorization approximation errors, and number of iterations

3.3.3 Computational complexity

The NMF, LNMF, NMFsc and ENMF were implemented using Matlab and experiments were conducted on a PC with 3.33GHz CPU and 4GB RAM. The CPU times, the number of iterations and the factorization approximation errors (in MSE) for learning the bases shown in Figure 3.1 are listed in Table 3.3. For fair comparison, the proposed ENMF stopping conditions with $\tau = 0.001$ and maximum number of iterations set to 7500 were used for all methods. Figure 3.4 shows the corresponding objective function vs. the number of iterations for each algorithm. (To generate the figure, algorithms were forced to keep running until reaching the maximum number of iterations.) It can be seen from Table. 3.3 and Figure 3.4 that ENMF converges much faster than LNMF and NMFsc. This is partially due to the adoption of the new initialization scheme and calculation of (instead of search for) step-sizes. Furthermore, the more localized basis images enable ENMF to achieve lower approximation errors compared to NMF and NMFsc.

As discussed above, besides the total number of training images $n$, image size $m$ and factorization rank $r$, the CPU time of the ENMF is also affected by the step size factor $\psi$ for updating $H$. The effect of different step size factors $\psi$ on the CPU time was tested using the BioID database with different number of images $n$, the results are shown in Figure 3.5. Notice that the ENMF converges faster with a larger $\psi$. However, it is found that the solution does not always converge when $\psi$ is more than 0.5. It is also interesting to see that the computational complexity of the ENMF grows linearly as the total number of training images increases.
3.3. Experimental results

Figure 3.4: Typical objective function vs. the number of iterations for NMF, LNMF, NMFsc and ENMF

Figure 3.5: The CPU times of ENMF learning with different values of $\psi$ and $n$

3.3.4 The discriminative power of ENMF

Following the standard PCA (eigenface) method, face recognition in a subspace can be performed as follows:

1. Feature extraction. After the subspace is learned from training images, let $\bar{t}$ be the mean of training images. Each training face image $t_i$ is projected into the learned subspace as a feature vector $f_i = W^{-1}(t_i - \bar{t})$ which is then used as a prototype feature point. A query face image $s$ to be classified is represented by its projection in the subspace as $f_s = W^{-1}(s - \bar{t})$.

2. Nearest neighbor classification. The Euclidean distance between the query and each prototype, $d(f_s, f_i)$, is calculated. The query is classified to the class to which the closest prototype belongs.
3.3. Experimental results

<table>
<thead>
<tr>
<th>Method</th>
<th>$r = 36$</th>
<th>$r = 49$</th>
<th>$r = 64$</th>
<th>$r = 81$</th>
<th>$r = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>0.43</td>
<td>0.41</td>
<td>0.38</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td>NMFsc</td>
<td>0.84</td>
<td>0.82</td>
<td>0.81</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>LNMF</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>ENMF</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 3.4: The recognition rates for different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>0.26</td>
</tr>
<tr>
<td>NMFsc</td>
<td>0.61</td>
</tr>
<tr>
<td>LNMF</td>
<td>0.64</td>
</tr>
<tr>
<td>ENMF</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 3.5: The recognition rates for different methods in single sample scenario

Based on the above simple scheme, the proposed ENMF can be directly employed for face recognition. We comparatively evaluated NMF, LNMF, NMFsc and ENMF representations for face recognition using the ORL database. In the evaluation experiment, all images from ORL database are directly used without normalization. The set of 10 images for each person is randomly partitioned into a training set of 5 images per person and a test set of the rest (the other 5 images per person). The training set is then used to learn the subspaces (basis images), and the test set for evaluation. All the compared methods take the same training and test data. For NMFsc, following Hoyer’s [50] work, we set $S_w$ (sparseness of the the basis matrix) to 0.75 and keep $S_h$ (sparseness of the the coefficient matrix) unconstrained. For ENMF, $S_h$ was set to 0.1, $\alpha$ was set to 1, $\psi$ was set to 0.5 and $\xi$ was set to 0.001. Table 3.3.4 shows the average recognition rates for the compared representations with different total number of bases $r$. As can be seen from the table, the performance of traditional NMF is surprisingly poor, ENMF achieves the best recognition accuracy.

The training set was then changed to containing only one image per person to test the four representations in the single sample scenario. We set $r$ to 81 for all the different methods and used all the rest images in the dataset as testing samples. The results are shown in Table 3.3.4. We see from the table that although ENMF still outperforms the other methods, face recognition purely based on ENMF coefficients suffer from the single sample problem, the recognition rate drops to 0.71 from 0.97.
3.4 Summary

NMF is extended in this chapter by imposing orthogonality constraint on the basis matrix and controlling the sparseness of the coefficient matrix for learning a compact local part-based representation of face images. Experimental results have demonstrated the efficiency of the proposed ENMF as a face representation. Semantically, the basis images obtained by ENMF imitate the abstracted local patterns suggested by Wallis et al., and can be regarded as the “middle level facial pattern” between pixels and actual facial parts. From pattern recognition point of view, localized facial features produced by ENMF offer advantages in face analysis, including stability to local deformations, lighting variations, and partial occlusion. As the proposed ENMF learning algorithm is a local minimizer, and gives different basis images from different initial conditions. A new initialization scheme has been developed to encourage the iterative learning process of ENMF to converge quickly to an instance of local part-based representation. In addition, the proposed $LC$ indicator provides an objective measure of the locality and compactness of an instance of local part-based representation.
Chapter 4

Modeling distinctiveness and familiarity

In the previous chapters of this thesis we alluded to two fundamental cognitive processes involved in human face perception: locating distinctive areas and measuring the familiarity of a face. These two processes can motivate computer vision-based face analysis techniques if appropriate computational models can be devised. In this chapter, reference face samples are proposed to represent previously encountered faces and computational models are developed for the two processes based on the ENMF face representation. The computational models are evaluated through experiments that follow the protocols commonly used in psychological studies and results are compared with subjective evaluation.

4.1 Finding distinctive facial regions

Studies in cognitive science have shown that humans tend to identify distinctive areas or regions of a new face by comparing the face with a general face model that is abstracted from the faces they have previously seen. We propose to establish a general face subspace by applying the ENMF to the entire reference face set and divide a face into either overlapping or non-overlapping regions. A general face model is constructed in the general face subspace with each face region being a basic unit for which the level of distinctiveness is to be measured. Due to the adoption of the ENMF subspace representation, the proposed computational model can be easily applied to regions of any shapes.
4.1. Finding distinctive facial regions

4.1.1 Construction of a general face model

Consider a set, \( R = \{ r_1, \ldots, r_m \} \), of reference faces consisting of \( m \) faces and each with size, \( n \) pixels. Let \( W \) be the ENMF subspace learned from \( R \), and \( H \) the coefficient matrix. Assume that a face is divided into \( K \) regions. We denote each region as \( z_k, k \in \{1, \ldots, K\} \). The general face model, \( G \), would consist of \( K \) appearance models; one for each region. The general face model is a set with \( K \) elements, i.e. \( G = \{ g_k \}, k = 1, \cdots, K \). The appearance model \( g_k \) is to be learned from the reference set for each region \( z_k \). Since the ENMF basis images are highly localized and compact, only limited bases contribute to a local facial region and each basis covers a limited facial regions. Therefore, the appearance of a facial region can be well described by a low dimensional feature vector, \( x \), formed by the coefficients in \( H \) that corresponds to the bases covering the region. For instance, for the region \( z_k \) of a face \( r_i \in R \), \( x \) is a subset of the \( i \)-th column in \( H, h_i \). It consists of the basis coefficients that correspond to the bases covering the region \( z_k \).

To model the knowledge abstracted from the reference set for each region \( z_k, k \in \{1, \ldots, K\} \), we approximate the distribution of the appearance features using a Gaussian mixture model (GMM) [110]. Such a model, with \( C \) components, can be written as

\[
p_k(x; \Theta_k) = \sum_{c=1}^{C} \alpha_c N(x; \mu_c, \Sigma_c), \ k \in \{1, \ldots, K\} \tag{4.1}
\]

where \( N(x; \mu_c, \Sigma_c) \) is a Gaussian function with mean value \( \mu_c \) and covariance matrix \( \Sigma_c \), corresponding to the \( c \)-th component. The positive mixing weights of the components, \( \alpha_c \), are chosen so that \( \sum_{c=1}^{C} \alpha_c = 1 \). Thus, \( \Theta_k = \{(\alpha_1, \mu_1, \Sigma_1), \ldots, (\alpha_C, \mu_C, \Sigma_C)\} \) represents the model parameters. \( p_k(x; \Theta_k) \) serves as the general model \( g_k \) for the region \( z_k \), whose parameters can be estimated from the corresponding reference set using expectation-maximization (EM) algorithm [110, 151].

4.1.2 Measurement of distinctiveness

We begin by denoting the distinctiveness of a region \( z_k, k \in \{1, \ldots, K\} \), on a given face \( s \), by \( d_k^s \). The face, \( s \), is projected onto the subspace \( W \) to obtain the corresponding coefficients, \( h \),
4.1. Finding distinctive facial regions

Figure 4.1: The normalization process

\[ h = W^{-1}s \] (4.2)

where \( W^{-1} \) is the pseudo inverse of the matrix \( W \). The corresponding appearance feature, \( x^k_s \), for the region \( z_k \) is formed by a subset of \( h \) that corresponds to the bases covering the region \( z_k \). The likelihood of \( x^k_s \) being realized from the model \( g_k \) is a good indicator of the distinctiveness, \( d^k_s \), of the region \( z_k \) on face \( s \) with respect to the reference face set \( R \). A high likelihood is indicative of a low distinctiveness. Hence, in our study, \( d^k_s \) is simply defined as the multiplicative inverse of the likelihood. This definition is intuitive as many applications require the relative degree of distinctiveness of regions on a face (i.e. ranking) based on a reference set. Hence the distinctiveness of a region \( z_k \) on a face \( s \) is defined as

\[ d^k_s \triangleq \frac{1}{p_k(x^k_s, \Theta_k)} \] (4.3)

4.1.3 Experimental results

All images from the BioID face database were employed as the reference set in the experiments. Face areas were first detected by the Viola-Jones [152] face detection method. A modified version of the Viola-Jones face detection method [180] was employed to find the areas of mouth and eyes within the detected face areas. Based on the detected positions of mouth and eyes, images were rotated so that the centers of eyes and mouth were at the same pixel coordinates for all images. The face areas were cropped and resized to \( 64 \times 64 \) face images. There are 1521 face images in total that were used as the reference set. Figure 4.1 illustrates the normalization process and sample face images are shown in Figure 4.2.

In the experiments, the ENMF was applied to all face images to learn the general face
4.1. Finding distinctive facial regions

Figure 4.2: Examples of normalized faces from BioID database

Figure 4.3: The ENMF bases learned from the BioID dataset

subspace. According to the study of ENMF in Chapter 3, the ENMF parameters were set to $r = 81, S_h = 0.1, \alpha = 1, \psi = 0.5$ and $\xi = 0.001$. Figure 4.3 shows the bases. Examples of original face images and their corresponding reconstructed images in the learned subspaces are shown in Figure 4.4.

Each face image was divided into rectangle local facial regions (each region being $16 \times 16$ pixels) with 4 pixel overlap in both horizontal and vertical directions. To train the Gaussian mixture model for each of the facial regions, the expectation maximization (EM) algorithm was used to estimate the parameters of the model [110, 151].

**Exaggerated samples** To show the efficiency of the proposed distinctiveness measure, 40 face images from ORL dataset, one image per subject, were used as test images. We first applied some random transforms on one local area of the test face images so that the facial region is clearly distorted. Then the proposed method was applied on the transformed
4.1. Finding distinctive facial regions

Figure 4.4: Examples of original face image (bottom row) and the corresponding reconstructed image (top row) in the learned subspace

Figure 4.5: Results for exaggerated faces

images to see if the distorted facial areas could be detected as the most distinctive regions. In all the cases, the method was able to identify the regions that cover the distorted areas as distinctive regions. Some typical examples are shown in Figure 4.5 in which the 10 most distinctive facial regions measured by the proposed method are marked in red squares. We can see that the distorted facial parts are well covered by the detected distinctive regions.

**Subjective test** A simple subjective test was also carried out to further evaluate how well the proposed method is in accordance with human subjective judgment. Forty (40) face images, one image per subject, were selected from the ORL database to form a test set and 20 participants were involved in the subjective evaluation experiments. Each participant was first presented with the face images from BioID dataset as the reference face set and then asked to click once on each test face image at which they considered as the center of the most distinctive region of the face. We then divided the faces into $16 \times 16$ regions with 4 pixel overlapping in both horizontal and vertical directions and employed the proposed method to measure the distinctiveness of these $16 \times 16$ facial regions for each image in the
4.1. Finding distinctive facial regions

Figure 4.6: An sample face image with the centers ("x") of distinctive areas labeled by the participants, the mean center ("+") and the most distinctive $16 \times 16$ region (the red rectangle) detected by the proposed algorithm.

test set. Based on the objective measurement, the facial regions of each face can be ranked according to the obtained distinctiveness.

Figure 4.6 show a typical example face image with the centers ("x") of distinctive areas labeled by the participants, the mean center ("+") and the most distinctive $16 \times 16$ region (the red rectangle) detected by the proposed algorithm. The overall performance of the objective ranking and subjective selection are summarized as follows.

- For any image, if the averaged position of the points clicked by all participants is considered as the true center of the distinctive region of the face, for 67.5% of the test images, the algorithm selected the most distinctive region that covers the true centers.

- For 85.0% of the test images, the algorithm selected the most distinctive region that covers at least one location clicked by the participants.

- For 83.5% of the test images, the top 5 distinctive regions selected by the algorithm cover the average position selected by all participants.
For 91.8% of test images, the top 10 distinctive regions selected by the algorithm cover the average position selected by all participants.

The above results indicate that the proposed distinctiveness measure matches the human subjective judgment reasonably well. The discrepancy is probably due to two major reasons. Firstly, participants would inevitably use the faces they have experienced before to construct their own general face model in the process of identifying the distinctive regions. In other words, the reference set used by the algorithm would eventually be different from the ones used by participants and the reference set for each participant would vary as well depending on the experience of the participant. Secondly, the algorithm only considered $16 \times 16$ regions with 4 pixel overlapping in both horizontal and vertical directions. However, participants did not have this restriction when they located the center of the region that was most distinctive to them.

### 4.2 Measuring face familiarity

The familiarity of a face is measured as the similarity between the face and a reference set that represents the previously seen faces. The exposure frequency is then modeled by the number of instances per subject in the reference set, the exposure intensity is modeled by the variations of the instances of each subject and the similar exposure is accounted for, by the instances of similar subjects in the reference set. Unlike the conventional similarity measure between a sample and a set of samples, the measurement of face familiarity not only has to account for the above three major factors that affect human perception of face familiarity, the measure also has to be able to take into account the two forms of familiarity, namely, the context-free and context-dependent familiarity.

#### 4.2.1 The proposed method

As suggested by Mandler, the key difference between context-free and context-dependent familiarity is whether specific previously encountered subjects contribute to the familiarity.
4.2. Measuring face familiarity

To address this difference and model both forms of familiarity based on the ENMF representation, the reference set is divided into a number of reference groups so that faces in each of the groups share certain similar properties in appearance. The familiarity of a given face against the reference set is measured by aggregating the similarities of the face to individual reference groups. Therefore, when faces in each reference group share certain common characteristics such as race, gender and age rather than personal features, the aggregated similarity models the context-free familiarity where no specific subjects make contributions. At the same time, when each reference group contains only faces of one specific individual, the context-dependent familiarity is modeled. To seamlessly model the two forms of familiarity, firstly, the reference groups need to be non-exclusive to each other and consist of both “common groups” and “personal groups”. Secondly, the manner of similarity aggregation should allow both kinds of reference groups contribute to the obtained familiarity so that “we are not able to distinguish between the two forms of familiarity”. [92].

In this thesis, an individual ENMF subspace is learned from each group of faces in the reference set. The similarity of a given face to a face group is then measured based on its reconstruction error after being projected into the learned ENMF subspaces. Aggregation is simply achieved by averaging the similarities obtained from the $\kappa$ most similar groups. The concept of the proposed familiarity measure is illustrated in Figure 4.7.

Let $R$ denote the reference set and $s$ a given sample face image. Further, let $C$ denote the total number of reference groups into which $R$ is divided and $R_t$ the reference groups indexed by $t \in \{1, 2, \ldots, C\}$. We also denote the similarity between $s$ and $R_t$ by $M(s, R_t)$. Finally, let $D_\kappa$ denote the set of indices of the $\kappa$ most similar reference groups to $s$. Then, the familiarity of the face $s$ with respect to reference set $R$ is defined as

$$F(s, R) = \sum_{t \in D_\kappa} \frac{M(s, R_t)}{\kappa},$$

where $\kappa \leq C$. Many methods can be adopted to measure the similarity between a sample and a dataset with different assumptions on the underlying statistical distribution of the dataset. According to the localized compact face representation described in Chapter 3, in this thesis,
4.2. Measuring face familiarity

Figure 4.7: The concept of the proposed familiarity measure
we measure the similarity $M(s, R_t)$ based on the reconstruction error after $s$ has been projected onto the ENMF subspace learned from $R_t$. Specifically, the coefficient vector, $h_t$, of $s$ in the ENMF subspace $W_t$ for $R_t$ can be obtained as

$$h_t = W_t^{-1} s,$$

(4.5)

where $W_t^{-1}$ is the pseudo inverse of matrix $W_t$. Based on the obtained coefficient vector $h_t$, the sample $s$ can be reconstructed by:

$$\hat{s} = W_t h_t.$$  

(4.6)

The reconstruction error between $s$ and $\hat{s}$ can be calculated by mean square error (MSE):

$$\text{MSE}(s, \hat{s}) = \frac{1}{n} E(s, \hat{s}) = \frac{1}{n} \sum_i (s_i - \hat{s}_i)^2,$$

(4.7)

where $n$ is the number of pixels in the face image. Since human perception has a logarithmic relationship to its stimuli, the similarity $M(s, R_t)$ is defined as the peak signal-to-noise ratio (PSNR) in order to mimic the non-linearity of human perception.

$$M(s, R_t) = \text{PSNR}(s, \hat{s}) = 10 \log_{10} \left( \frac{\text{MAX}_I}{\text{MSE}(s, \hat{s})} \right),$$

(4.8)

where $\text{MAX}_I$ is the maximum possible pixel value in the image. PSNR is widely used in image/video compression to measure the quality of an image or video with respect to a reference image or video. It is one of the commonly used measurements that are believed to be in a reasonable agreement with human perception.

Note that in the absence of label information (identity, age, race, gender, etc) in the reference set, we are still able to divide the reference set into “common groups” and “personal groups”. One simple but effective way is to employ clustering techniques with a controlled number of clusters. When the reference set is clustered into small number of groups, ENMF subspaces learned from these groups correspond to common features of many subjects, thus the clustered groups can be used as “common groups”. Similarly, when the reference set is clustered into a large number of groups, these groups can be treated as “personal groups”,
since the learned ENMF subspaces from the groups potentially correspond to specific personal features. In addition, the reference set can also be clustered into different groups using different features.

The choice of $\kappa$ can be guided by the intuition that humans usually recall only a small set of most familiar faces during perception and as such should be relatively small. By choosing an appropriately small value of $\kappa$, the selected most similar reference groups contribute to the familiarity. These selected groups may be all specific “personal groups” or all “common groups” that involve many subjects or, with high possibility, a mixture of both kinds of groups depending on the contributions of the two forms of familiarity.

### 4.2.2 Experimental results

Extensive experiments were conducted to evaluate the effectiveness of the proposed method for familiarity measurement. In the experiments, we simulated the face familiarity ranking task that is commonly used in psychological studies. The familiarity of known faces, prototype faces, and unknown faces were measured objectively by the proposed method and subjectively by humans. The ORL and BioId face databases were used in the experiments. All face images are normalized as described in Section 4.1.3. For ENMF parameter settings, we still chose $r = 81$, and set $r = 81, S_h = 0.1, \alpha = 1, \psi = 0.5, \xi = 0.001$.

Faces of individuals in the ORL database were used as known faces and five different reference sets were created from the database to simulate different levels of exposure frequency and intensity. Each of the reference sets was created by randomly selecting a number, $\zeta$, of face images per subject from the ORL database; we chose $\zeta = 10, 8, 5, 2, 1$. When $\zeta = 10$, the known faces for testing appeared in the reference set since ORL only has 10 images per subject. For other values of $\zeta$, the known face for testing may not necessarily appear in the reference set. Three different partition schemes were used to divide each reference set into common and personal groups:

**Common groups.** k-means clustering [44] was employed to partition the database into different numbers of groups.
4.2. Measuring face familiarity

Personal groups. Reference groups were created based on identity information, each group corresponds to one subject in the ORL database and only contains faces of the subject.

Combined groups. The reference groups consist of all the personal groups and all the clustered groups at $C = 1, 5, 10, 20, 50$.

Morphs generated based on known faces were used as prototype faces for testing. Specifically, 25 pairs of face images were selected from ORL database as parent faces. For each pair of parent faces, a morph was generated with an equal contribution from each parent face. All the morph images were generated nonlinearly using morphing software FantaMorph4 [52] based on manually labeled facial landmarks. Some morphs and their parent faces are shown in Figure 4.8.

During the test, the degree of familiarity of all the morphs, 50 known faces selected from ORL database, and 50 unknown faces selected from BioID database were measured by the proposed method using the five reference sets. Figure 4.9 shows an example of the calculated familiarity of a morph, a known parent face of the morph, and an unknown face and their ten most similar subjects in the reference set being divided into personal groups (the subjects are listed left to right from the most similar to the least).

Context-free familiarity The proposed method was first tested in the context-free scenario, using clustered reference groups with small numbers of $C$ ($C = 1, 5, 10, 20$) and different values of $\kappa$ (ranging from 1 to 8, depending on the value of $C$). While absolute values
4.2. Measuring face familiarity

Figure 4.9: The familiarity of three sample faces ($\kappa = 10, \zeta = 10$ using personal reference groups) and the faces of their ten most similar subjects in the reference set (left to right, from the most similar to the least)

of the measured familiarity varies with the change of $C$ and $\kappa$, the relative values of the familiarity among the test faces turned out to be relatively stable. Therefore, only representative cases are reported in this section. In addition, the measurements (in PSNR) are normalized to the range of (0, 1) for easy comparisons (especially with the subjective test to be presented later).

Figure 4.10 shows the average familiarity of known faces, morphs, and unknown faces measured when clustering the reference set into one group, i.e. $C = 1$. As seen from the figure, three kinds of faces are clearly separated. Unknown faces obtain the lowest familiarity and the unknown morphs generated, based on known faces are measured as quite more familiar than the actual known ones. This is mainly because, in the case $C = 1$, the whole reference set is used as a single reference group. The proposed method models the context-free familiarity based on common features of all the faces in the reference set, and each known face contributes to the obtained familiarity. Therefore, faces similar to more than one known subject are measured to be considerably more familiar than known faces themselves; a high level of prototype effect is exhibited. This result is in line with the Mandler’s theory that context-free familiarity causes high level of prototype effect. However, since all the known faces contribute to the familiarity, the influence of factors of exposure frequency, exposure intensity and similar exposures is weakened. We can see that the increase of familiarity for all the three kinds of faces is not distinct as the number of face images per subject in the reference set (the value of $\zeta$) increases. This situation is improved when using more clustered reference groups and choosing an appropriately small value of $\kappa$, so that only a number of
4.2. Measuring face familiarity

Figure 4.10: Results for using clustered reference groups in the case of $C = 1$

Figure 4.11: Results for using clustered reference groups in the case: $C = 10, \kappa = 3$

selected known faces that share certain common features with the given test face contribute to the familiarity measure. Figure 4.11 shows the results for the case of $C = 10, \kappa = 3$. We can see that compared with the case of $C = 1$, as the level of exposure intensity and frequency (the value of $\zeta$) increases, the increase in the value of familiarity obtained is more significant. The familiarity increase for unknown faces accounts for the the factor of “similar exposure”; that is, prior exposure to the faces that belong to different persons also contributes to the familiarity of a face of unknown person.

**Context-dependent familiarity** We now report on the evaluation of the proposed method when context-dependent familiarity is modeled. It has been observed that the influence of exposure frequency and intensity upon the familiarity in these cases is similar to that in the context-free case. We, therefore, present here how the familiarity varies against the value of $\kappa$ by setting $\zeta = 10$. Figure 4.12 shows the normalized average familiarity of three types of test faces measured by the proposed method using personal reference groups. We can see from the figure that unknown faces always obtain low level of familiarity for all the $\kappa$. When $\kappa = 1$, known faces are measured as more familiar than the morphs. In this case, a given face is matched to the most similar known subject (group). A known face would always find
4.2. Measuring face familiarity

Figure 4.12: Results for using personal reference groups when $\zeta = 10$

a good match (reconstruction), while the morph could only obtain a partial match. Thus no prototype effect is shown for $\kappa = 1$. This corresponds to the situation when the recognition of a known face occurs, where we are able to specifically recall the particular encounter. When $\kappa = 2$, a morph could be partially matched to both its parent subjects, and a known face could still find only one good match. Therefore, as can be seen from the figures at $\kappa = 2$, the average familiarity for known faces and morphs are similar, the prototype effect begins to appear. As the value of $\kappa$ further increases, more known subjects contribute to the matching, until it reaches the maximum value, the familiarity is measured based on all the known faces just like the case of $C = 1$. So we can see for large values of $\kappa$, similar results to the ones shown in Figure 4.10 are obtained; a higher level of prototype effect is exhibited.

As discussed in Section 4.2.1, without using the identity information we can still create “approximate personal groups” by clustering the reference set into a large number of groups. In this experiment, we approximate the personal reference groups by clustering the reference set into 50 groups ($C = 50$), which is a relatively large number for the ORL database that only contains 400 images of 40 subjects. Figure 4.13 shows the average familiarity measured by using the approximate personal groups. As expected, the measured familiarity are comparable to those of using personal reference groups shown in Figure 4.12, where the prototype effect begins to show at $\kappa = 2$.

**Overall familiarity** According to results from cognitive studies, the two forms of face familiarity, namely, context-free and context-dependent, happen simultaneously and are usually not distinguishable in the perception process of face familiarity. The overall familiarity of a face may be contributed by both forms, therefore, it should be measured by using the
4.2. Measuring face familiarity

Figure 4.13: Results for using clustered reference groups with $C = 50, \zeta = 10$

combined reference groups that are composed of personal groups and all clustered groups at different $C$ values. In this way, the two forms of familiarity and the three factors that affect face familiarity are modeled simultaneously by the proposed method. Figure 4.16 shows the results at $\kappa = 3$ (error bars are shown here for comparison with the subjective test to be presented later). We can see from the figure that a clear prototype effect is exhibited and as the value of $\zeta$ increases, higher degrees of familiarity are obtained for all the test faces. The familiarity increase for morphs and known faces is limited when the value of $\zeta$ exceeds 5. This scenario is in line with our experience that face familiarity is gained easily and quickly during first few encounters, while the later process when the face becomes more familiar to us is relatively slow.

To further verify that the two forms of face familiarity are not distinguishable, Table 4.1 shows how each type of reference groups contributes to the obtained familiarity. For the three types of test faces (known parent faces, morphs, unknown faces), we calculate the percentages of the types of the reference groups selected as the most, second most and third most similar groups to the faces. When considering the clustered reference groups at $C = 50$ as personal groups and all other clustered groups as common groups, the table has shown that the familiarity of unknown faces is mainly contributed by the common groups. As for the known faces and morphs, the obtained familiarity is measured based on a mixture of common groups and personal groups, which demonstrates the proposed method has effectively integrated the context-free and context-dependent familiarity.

To validate how the objective measurement conforms to human perception, a subjective study of the familiarity ranking task was conducted using exactly the same set of training
4.3. Summary

This chapter proposes to use a set of reference samples to model the previously encountered faces. Based on the ENMF representation, two models are developed to measure the distinctiveness of facial areas and familiarity of a face with respect to the reference set, respectively.

<table>
<thead>
<tr>
<th>Group</th>
<th>Known faces</th>
<th>Morphs</th>
<th>Unknown faces</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
</tr>
<tr>
<td>Personal</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>c=1</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>c=10</td>
<td>0%</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>c=20</td>
<td>0%</td>
<td>21%</td>
<td>34%</td>
</tr>
<tr>
<td>c=50</td>
<td>0%</td>
<td>53%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Table 4.1: The percentages of each type of reference groups selected as the most, second most, third most similar group to the testing faces.

and test data as was used for the objective measurement. The study involves two phases: training and testing. During the training phase, 50 participants were asked to study and remember the face images from ORL database. The 50 participants were divided into 5 equal groups. For each group, face images in one of the five reference sets used in the objective measurement ($\zeta = 10, 8, 5, 2, 1$ respectively) were displayed to the participants. There was no time limitation for the participants to study each subject, images of a different subject were displayed after the “next” button on training interface was clicked, and each subject was displayed twice. A snapshot of the training interface ($\zeta = 10$) is shown in Figure 4.14. In the testing phase, participants were asked to rate the familiarity of given test face images. Five point Likert-type scale were used to measure the familiarity subjectively: very familiar (5), familiar (4), neutral (3), unfamiliar (2), totally unknown (1). A snapshot of the testing interface in subjective study is shown in Figure 4.15. Figure 4.17 shows the results of normalized subjective ranking. Comparing the subjective familiarity (Figure 4.17) with the objective measurements (Figure 4.16), it can be seen that the proposed objective measure is highly consistent with the subjective ranking.
4.3. Summary

Figure 4.14: A snapshot of the training interface in subjective study

Figure 4.15: A snapshot of the testing interface in subjective study

Figure 4.16: Results for using combined reference groups with $\kappa = 3$

Figure 4.17: Results of the subjective ranking with error bars
Rather than focusing on the distinctiveness of whole face, as individual facial regions play the key role to make one face different from another, the proposed distinctiveness model measures the distinctiveness of each facial area and thus identify the most discriminative ones. In the proposed familiarity model, the exposure frequency is modeled by the number of instances per subject in the reference set, the exposure intensity is modeled by the variations of the instances of each subject and the similar exposure is accounted for, by the instances of similar subjects in the reference set. The distinctiveness and familiarity measures were evaluated through experiments that follow the protocols commonly used in psychological study and are compared with subjective evaluation. Results have shown that the proposed measures are highly consistent with the subjective judgments.
In this chapter, novel methods are proposed to address the single sample face recognition problem based on the computational model for distinctiveness. As reviewed in Chapter 2, identifying the most distinctive facial areas is a key process in human face perception. However, this process has not been fully exploited in any of the existing face recognition methods. In holistic-based methods, the process can hardly be incorporated since they use global information from the entire face and implicitly treat all facial areas as equally important. In the local-based methods that extract features from local facial regions, the common approach either makes the assumption that anatomical facial components are the most useful areas for recognition or locates the most salient parts within the face and treats the salient parts equally. The proposed measure of distinctiveness offers an effective way to simultaneously select and/or weight facial regions adaptively. Therefore, this chapter explores how the distinctiveness measure can improve local-based methods. In addition, a hierarchical method is proposed to effectively incorporate the configural information for recognition.

5.1 Overview of local-based methods

In general, a local-based method, as shown in Figure 5.1, consists of four key steps: defining local regions, extracting local features, weighting local regions, matching and fusing. The commonly used approaches to defining local regions are either partitioning of the face into equal areas (e.g. as shown Figure 5.2 (a)), or extraction of local patches at specific locations (e.g. as shown Figure 5.2 (b)). For the former, the partitioned regions may have various
shapes and sizes and are either overlapped or non-overlapped with each other [141, 2, 18]. In the case of the latter, different methods have been proposed to select the local facial regions. Earlier methods select regions around predefined fiducial points (facial landmarks) such as eye corners, mouth corners, eyebrow corners and nostril corners. Recently, detectors of interesting points have been employed to select salient facial regions. For example, scale invariant feature transform (SIFT) [88] detects key points in an image by means of a local optimization process applied to the differences of Gaussian images at different scales and orientations. Unlike the SIFT descriptor which tends to look for points having blob-like features, the Harris-Laplace detector [99] searches for points in the image whose value of cornerness is locally maximal and was employed to select local regions for face recognition in [29]. In addition, Kepenekci et al. [61] detected interesting points based on Gabor features under the assumption that points with high-energized Gabor wavelet response contain more information about a face image.

Once the local regions are defined or selected, appropriate feature is extracted from each facial region to represent the local information. The commonly used features include gray-value feature [96, 141, 18], and Gabor feature [158, 61]. Recently, other derived features such as LBP [2], SIFT [11, 90] and Haar features [113] have also been studied in face recognition.

Because the useful information for recognition is usually not uniformly distributed on a face image, the various facial regions should not be treated as equally important and, thus,
5.1. Overview of local-based methods

weights are usually assigned to each region based on its importance. For equally partitioned local regions, weights are often predefined based on the assumption that some specific anatomical facial components (in most cases the eyes and nose) are more discriminative than others. For example in [2], as shown in Figure 5.3, eyes are given the highest weight while cheek regions are given zero weight. However, the assumption is not always true as characteristic facial features vary from one individual to another and they cannot always be on eyes and nose. For some individuals, other facial features such as eyebrows, cheek, chins and the gaps around the eyes, nose and mouth could be more discriminative for recognition, especially when they contain scars, spots, dimples and lines. As a result, predefined weights would introduce adverse effect when high weights are assigned to non-discriminative facial regions. Such an adverse effect is, to some extent mitigated by selecting local regions at interest points and treating the regions equally important, i.e. assigning the same weight to all selected regions. However, the interest points are not necessarily aligned with the visual discriminative regions for recognition and mismatch of the interest points between the test face and sample face would also introduce challenges in the recognition task.

In the final step, corresponding local regions between the sample and test face are matched and the region-based matching scores are fused into an overall matching score from which a recognition decision is made. Assume $n$ local facial regions are considered, let $d_i$ denote the
5.2 Improved local-based methods

We propose to improve the local-based methods by weighting (or selecting) facial regions adaptively according to their distinctiveness. Figure 5.4 shows the block diagram. Specifically, distinctive facial regions are identified from the sample images with respect to a large set of reference face samples. For each sample image (i.e. each subject to be recognized), higher weights are assigned to more distinctive regions than less distinctiveness areas and

where $d_i$ is obtained by comparing the feature vectors $F_i^t$ and $F_i^s$, which are extracted from the region $G_i$ in the test face image $t$ and sample face image $s$ respectively. Note that

$$d_i = d(F_i^t, F_i^s)$$

where $d()$ is the similarity (dissimilarity) measure.

**5.2 Improved local-based methods**

Figure 5.3: Facial region partition (a) and weights (b) for the local regions in [2]. The facial image is equally divided into non-overlapped $7\times7$ blocks. The black squares indicate weight 0.0, dark gray 1.0, light gray 2.0 and white 4.0

The overall matching score is often defined as a linear integration of the matching scores obtained from each region:

$$D = \sum_i W_i d_i$$

(5.1)

where $d_i$ is obtained by comparing the feature vectors $F_i^t$ and $F_i^s$, which are extracted from the region $G_i$ in the test face image $t$ and sample face image $s$ respectively. Note that

$$d_i = d(F_i^t, F_i^s)$$

(5.2)
5.2. Improved local-based methods

Figure 5.4: The block diagram of improved local-based methods inspired by distinctiveness measure

thus, the weights are adaptive to individuals. To verify the efficiency, we employ the proposed weighting scheme in two representative local-based methods. Specifically, we selected the work of Ahonen et al. [2] in which the face areas are equally partitioned and that of Kepenekci et al.’s work [61] in which the selected regions at interest points.

5.2.1 Improved LBP-based method

Ahonen et al. [2] equally partitioned a face image into non-overlapped $7 \times 7$ local regions, as shown in Figure 5.3, and extracted the local binary pattern (LBP) histogram [106] as feature. The weight $W_i$ for each region $G_i$ was predefined as shown in the same figure, where the black squares indicate weight 0.0, dark gray 1.0, light gray 2.0 and white 4.0. Several distance metrics were used as the dissimilarity measure $d()$, including histogram intersection, log-likelihood statistic and Chi square statistic. Due to the stability and simplicity of the LBP-based face representation, this method has attracted much attention in the past. We measure the distinctiveness of each non-overlapped $7 \times 7$ facial region for the sample face and assign high weights to distinctive regions. That is,

$$W_i = \frac{c_i}{\sum_j c_j} \quad (5.3)$$

where $c_i$ is the distinctiveness for region $G_i$. 
5.2. Improved local-based methods

5.2.2 Improved Gabor-based method

In the work reported by Kepenekci et al. [61] (referred to as selected Gabor method in the rest of the thesis), 40 Gabor wavelets were convolved with a face image and interesting points were detected at the locations with high-energized Gabor wavelet response on each of the 40 Gabor filtered images. At each interest point, Gabor features were extracted. Let \((x'_j, y'_j), F'_j, (j = 1, 2, \cdots, N')\) be the locations and feature vectors of the \(N'\) interest points of a test image and \((x'_k, y'_k), F'_k, (k = 1, 2, \cdots, N')\) be the locations and feature vectors of the \(N'\) interest points of a sample image. The matching between the test image and sample image goes through the following process. For each interest point \(j, j = 1, 2, \cdots, N'\), in the test image, search for the possible matching interest point \(k, k = 1, 2, \cdots, N'\), in the sample image, that satisfies the following two criteria:

1. 

\[
\sqrt{(x'_k - x'_j)^2 + (y'_k - y'_j)^2} < th_1
\]

(5.4)

where \(th_1\) is a threshold usually set to the approximate radius of the region of facial components.

2. 

\[d(F'_j, F'_k) < th_2\]

(5.5)

where \(th_2\) is learned from the gallery sample set.

If no such an interest point is found in the sample image, then the interest point \(j\) in the test image is ignored in the matching. The overall matching score or the integrated distance \(D\) of the two faces is calculated as:

\[
D = \frac{\sum dis_j}{n'}
\]

(5.6)

where \(n'\) is the total number of interest points in the test image that have matching interest points in the sample image. Assume there are \(n'_j\) interest points in the sample image found to match the interest point \(j\) in the test image, \(dis_j\) is defined as:

\[
dis_j = \min\{d(F'_j, F'_1), d(F'_j, F'_2), \ldots, d(F'_j, F'_{n'_j})\}
\]

(5.7)
5.3. A hierarchical method

where $d(F_j^r, F_k^r), k = 1, 2, \cdots, n^r_j$, is the Euclidean distance between $F_j^r$ and $F_k^r$.

The selected Gabor method was modified by locating the interest points in the test and sample images at the centers of the most distinctive regions. To this end, face images were partitioned into $8 \times 8$ regions with 50% overlaps, and the distinctiveness of each region was measured. Then, the centres of the $N$ most distinctive regions were selected as the interest points and Gabor features were extracted at each of the interest points.

5.2.3 Experimental results

Both the original LBP and the selected Gabor methods were evaluated on the FERET database using the standard FERET evaluation protocol [119] which sets one training sample per subject. To verify the efficiency, the improved LBP and Gabor-based methods were also tested on the FERET database using the same protocol [119]. The protocol has the gallery set consisting of 1196 frontal images of 1196 persons and four probe sets: $\text{fabf}$ (1195 images with expression variations); $\text{fabc}$ (194 images with illumination variations); $\text{dup.I}$ (722 images taken in less than 18 months); $\text{dup.II}$ (234 images taken about 18 months later). In the experiments, only Chi square statistic was used as the dissimilarity measure $d()$ for the original and improved LBP method. The parameters for the original and improved Gabor were set as follows: $th_1 = 15$ pixels, $th_2$ was set to the mean Euclidean distance of all pairs of Gabor feature vectors in the gallery set and $N = 100$.

The average recognition rates of the original and improved LBP and selected Gabor methods are showed in Table 5.1. For the sake of comparison, reported results of two SIFT feature-based methods, SIFT_Gird [11] and SIFT_Person-specific [90] on the same FERET database are also included in the table. It can be seen that the recognition rates for the modified LBP and Gabor methods have been noticeably improved.

5.3 A hierarchical method

Besides identifying the most discriminative facial areas, another feasible way to improve local based methods is to effectively incorporate the spatial/configurational information in faces.
5.3. A hierarchical method

<table>
<thead>
<tr>
<th>Method</th>
<th>fabb</th>
<th>fabc</th>
<th>dup.I</th>
<th>dup.II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected Gabor [61]</td>
<td>0.96</td>
<td>0.70</td>
<td>0.58</td>
<td>0.47</td>
</tr>
<tr>
<td>Modified selected Gabor</td>
<td>0.97</td>
<td>0.71</td>
<td>0.64</td>
<td>0.59</td>
</tr>
<tr>
<td>LBP [2]</td>
<td>0.97</td>
<td>0.79</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>Modified LBP</td>
<td>0.97</td>
<td>0.79</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>SIFT_GRID [11]</td>
<td>0.94</td>
<td>0.35</td>
<td>0.53</td>
<td>0.36</td>
</tr>
<tr>
<td>SIFT_Person-specific [90]</td>
<td>0.97</td>
<td>0.47</td>
<td>0.61</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 5.1: The recognition rates of different methods for the standard FERET evaluation protocols

Existing local methods consider the spatial information through Equation 5.1, where similarity scores of local feature vectors extracted from the same image region are integrated to obtain a global score for final decision. However, since facial appearance on 2D images is sensitive to viewpoint variations, the same image region of different face images may not correspond to the same facial area even when faces are aligned, and the correspondency of local feature vectors is not able to be ensured especially for small image regions. In other words, the spacial information within small image regions is meaningless and often harmful for face recognition. Therefore, in this section, a novel local method is proposed based on hierarchical partition and representation of face images. This strategy allows only the global configural structure of faces to be considered while the detail spatial information within individual facial regions is ignored. Figure 5.5 shows the block diagram of the proposed hierarchical method. The method consists of three key components: hierarchical face partitioning, regional face representation, matching and weighted fusion.

5.3.1 Hierarchical face partition

We begin by dividing a face image into $N$ regions. Each region $G_i (i = 1, 2, \cdots, N)$ is then further divided into $M_i$ sub-regions. For each sub-region $g_{k,i} (k = 1, 2, \cdots, M_i; i = 1, 2, \cdots, N)$, a feature vector $f_{k,i}$ is extracted. As we do not consider spatial information within each $G_i$, defining $G_i$ is critical in the proposed method. The choice of a large $G_i$ will result in the loss of spatial information, while a small $G_i$ results in sensitivity to viewpoint variations. We
propose to partition the face meaningfully and let each $G_i$ contain one and only one key facial component, such as nose, eye and eye brow. Thus the detail spatial relationship between feature vectors extracted from the same facial component is ignored and the global configural structure of faces is retained and emphasized. Note that since face image alignment and normalization are necessary steps in most face recognition methods [186] (to guarantee that key facial landmarks such as eyes and mouth centers are at the same image coordinate for all face images), no extra processes are involved for the facial component-based face partition. Figure 5.6 shows an example of a hierarchically partitioned face image.

5.3.2 Regional face representation

In order to represent each facial region $G_i$ using feature vectors $f_{ki}(k = 1, 2, \cdots, M_i)$ without considering their spacial relationships, the regional representation for $G_i$ should focus on the distribution of the local feature vectors rather than the individual features themselves. To this end, in the proposed method, each feature vector $f_{ki}$ is first quantized according to
5.3. A hierarchical method

Figure 5.6: A hierarchically partitioned face image. The face is divided into 6 regions $G_i$ (red blocks), each $G_i$ is further divided into sub-regions $g_{k,i}$ (blue blocks).

A codebook. The codebook is constructed using clustering approach based on a large set of reference face samples. Thus, each code vector in the codebook represents a kind of local facial pattern that frequently appears in human face images. With a limited codebook size, the learned code vectors generally tend to only encode inter-person variations, since the number of images from different subjects in the reference set is far more than that from the same subject. Therefore, ideally, when face images are projected into the quantized feature space, different face images from the same individual would converge around a particular location in the space. Then, each facial region $G_i$ is represented by a regional histogram $F_i$, that is obtained by counting the occurrence of the quantized feature vectors within $G_i$. Formally:

$$F_{j,i} = \sum_k I\{Q(f_{k,i}) = j\}, j = 1, \ldots, q, k = 1, 2, \ldots, M_i$$ (5.8)

Where $Q()$ is the quantization process mapping $f_{k,i}$ into the learned quantized feature space, $j$ is the index of code vectors, $q$ is the size of codebook, $k$ is the index of $G_i$’s sub-regions, $M_i$ is the total number of $G_i$’s sub-regions, and

$$I\{A\} = \begin{cases} 1, & \text{if } n A \text{ is true} \\ 0, & \text{if } n A \text{ is false} \end{cases}$$
5.3.3 Matching and weighted fusion

The facial regions $G_i$ are weighted according to the distinctiveness of corresponding facial component for the sample face. Many studies [121] have shown that due to shape variations caused by expressions, the mouth is not considered as good as the upper-face features. Thus, the distinctiveness of the mouth is not measured in the proposed method. When weights are assigned to the facial regions, the lowest weight is always given to the mouth part and other facial regions are assigned weights according to the measure of distinctiveness:

$$W_i = \frac{c_i}{\sum_i c_i + \min(c_i)}$$  \hspace{1cm} (5.9)

In Equation (5.9), $i$ is the index of different facial regions except mouth part and $c_i$ is the distinctiveness of different facial regions. For the mouth region,

$$W_m = \min(w_i)$$

The Chi square statistic can be used as the similarity measure $d()$ to calculate the regional matching score $d_i$:

$$d_i = d(F^s_i, F^t_i) = \sum_j \frac{(s_{ji} - t_{ji})^2}{s_{ji} + t_{ji}}$$  \hspace{1cm} (5.10)

In Equation (5.10) $s_{ji}$ is the frequency of a code vector $j$ that belongs to $G_i$ of a sample image $s$. Similarly, $t_{ji}$ is the frequency of a code vector belonging to test image $t$.

Finally, test images are compared with the sample image and the best match is output as recognition result by searching the best overall matching score $D$ defined in Equation 5.1.

5.3.4 Experimental results

In the experiments, we implement the proposed hierarchical method using three different local features. First, the details of experimental setup is described with respect to facial region partitioning, local feature extraction and codebook generation. Then, we present the experiments conducted based on ORL face database for investigating the effect of different parameters. Finally, the proposed method is evaluated based on the standard FERET evaluation protocol.
5.3. A hierarchical method

Experimental setup

Facial region partition. All the faces were first normalized into $64 \times 64$ as described in Chapter 4. The normalized face was divided into 6 regions based on facial components as shown in Figure 5.6. Note that the eyebrows were divided separately, due to the study from O’Toole et al. [108], who claim that of the different facial components, eyebrows are among the most important for recognition. Then the facial region was further divided into $8 \times 8$ pixel sub-blocks with overlaps. The degree of overlap affects the recognition performance and we will evaluate the effects of varying sub-block overlap later.

Local feature extraction. Three different local features were used in the experiment. For the first feature we calculate the 2D-DCT in each sub-block to extract the feature vector $f_{k,i}$ [148]; this results in a $8 \times 8$ DCT coefficient matrix with 64 coefficients. Only a subset of the DCT coefficients were used to represent each sub-block. Ideally, the subset of coefficients should be insensitive to illumination variations and contain as few elements as possible while representing all the necessary information (the contributions of different subsets to the recognition performance is investigated later).

The second feature is the LBP feature [106, 107]. The LBP operator forms labels for image pixels by thresholding the neighborhood with the center value and representing the result as a binary number. The histogram of the labels of the pixels within each sub-block can then be used as the feature vector $f_{k,i}$.

The last feature is the ENMF representation introduced in Chapter 4. An ENMF subspace is learned from the reference set and a sample face is represented by its coefficient vector in the learned subspace. A sub-block of the sample face can be represented by a subset of the the coefficient vector that corresponds to the bases covering the sub-region.

Codebook generation. The BioID face database was used as the reference set. The codebook for all the sub-blocks was generated using “k-means clustering” [44] from the reference set. To capture all possible inter-person variations, the training datasets should be large enough and allow for diverse variation. At the same time, the size of codebook cannot be too small. On the other hand, to avoid code vectors that only capture intra-person
5.3. A hierarchical method

Figure 5.7: Examples of original face image and the corresponding reconstructed images (DCT features were used)

variations, the codebook size cannot be too large. The effects of different codebook sizes are investigated later in this chapter. After the codebook has been trained, all the sub-blocks from the face are mapped to the code vectors in quantized feature space using a nearest neighbor strategy; the regional histograms are also calculated. Figure 5.7 shows examples of original face images and their corresponding reconstructed images (DCT features were used). Figure 5.8 and Figure 5.9 show typical examples of regional histograms of left eye based on DCT features and ENMF features respectively. As can be seen, histograms of different persons are clearly different while histograms of the same person show some similarity, though there is a small difference in detail.

Parameter Selection

Experiments were conducted, using the ORL face database, to investigate the effect of different parameters. For each test, one image was selected as the training sample while all the other images in the database were used as testing samples.

Using DCT features. The DCT coefficients were scanned in a zig-zag pattern (so-called Peano scanning) [57] to obtain an ordered list of coefficients in which the first few coefficients corresponding to low-frequency components contain the most important information. In particular, the first DCT coefficient (DC component) reflects the average pixel value inside each sub-block and hence would be affected the most by any illumination change. The second and third coefficients ((0, 1), (1, 0) component) represent the average horizontal and
5.3. A hierarchical method

Figure 5.8: Typical examples of regional histograms of left eye region (DCT features were used)

Figure 5.9: Typical examples of regional histograms of left eye region (ENMF features were used)
vertical pixel intensity change, respectively. As such, they would also be significantly affected by illumination changes. Thus, to find the optimal subset mentioned above, following DCT subsets were examined in the experiment:

- **S1**, the first 16 coefficients.
- **S2**, removing the DC component, and selecting the next 16 coefficients in the ordered list.
- **S3**, removing the DC, (0, 1) and (1, 0) components, and selecting the next 16 coefficients in the ordered list.
- **S4**, the first 8 coefficients.
- **S5**, removing the DC component, and selecting the next 8 coefficients in the ordered list.
- **S6**, removing the DC, (0, 1) and (1, 0) components, and selecting the next 8 coefficients in the ordered list.

In the experiment, for different DCT features, corresponding codebook with fixed size of 64 were generated and used. To ensure adequate representation of the face, face images were divided by sliding the 8 × 8 partition template one pixel at a time, thus, sub-blocks with 63/64 overlap were used. The recognition performance for different DCT subsets is shown in Figure 5.10. As can be observed, DCT subsets with less elements achieved better results. This is probably due to the fact that most of the information is already captured and subsequent coefficients add “noise” to the representation. By discarding the elements affected by illumination variations, the recognition rate is improved. The best results are obtained by using subset S5, which only removes the DC component.

One direct effect of varying the overlap of sub-block is that the number of feature vectors extracted from a face image grows in an exponential manner as the overlap is increased. Experiment is conducted here to investigate the indirect effect of overlaps on the final recognition performance. In the experiment, DCT subset S5 was adopted for feature extraction
and the corresponding codebook with 64 code vectors was used for quantization. The experimental results are shown in Figure 5.11. As can be seen from the curve, by increasing the percentage of overlap, the recognition performance is improved. This may be explained as that the code vector distribution for one image can be better estimated based on more extracted features. However, we can see the increase of recognition rate is limited when the overlap exceeds 75%.

To find the most appropriate codebook size, in the experiment, 4 different codebooks were generated based on the same training data but only varying the codebook size. These codebooks were tested while keeping the other two parameter unchanged (DCT subset S5 and 63/64 sub-block overlap was used); the result is shown in Figure 5.12. As we analyzed
Figure 5.12: Effect of different codebook sizes (DCT feature)

before, codebooks with middle range of the size achieved better recognition rates.

**Using LBP features.** By using a circular neighborhood and bilinear interpolation at non-integer pixel coordinates, the LBP operator can be defined for any radius and number of sampling points. Three different LBP operators i.e. \( LBP_{(8,1)} \), \( LBP_{(8,2)} \) and \( LBP_{(16,2)} \) were first tested. \( LBP_{(P,R)} \) denote the LBP operator with \( P \) sampling points on a circle of radius of \( R \) pixels). In these experiments, codebook size of 64 and sub-block overlap of 50% were used. Furthermore, following previous works [2, 107], the idea of uniform patterns was used so that the LBP histogram for each sub-block has a separate bin for every uniform pattern and all nonuniform patterns were assigned to a single bin. An LBP code is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular. The average recognition rate achieved by using \( LBP_{(8,1)} \), \( LBP_{(8,2)} \) and \( LBP_{(16,2)} \) are 0.82, 0.88 and 0.92 respectively. \( LBP_{(16,2)} \) was selected and tested with different codebook sizes with sub-block overlap of 50%. The results are shown in Figure 5.13. Finally, the effect of varying the overlap of sub-block was tested using \( LBP_{(16,2)} \) and codebook size of 64, the result is shown in Figure 5.14. As seen from the figures, the best result was achieved with codebook size of 64 and overlap of 75%. Changes in sub-block overlaps did not affect the performance of LBP features significantly.

**Using ENMF features.** Similar experiments were conducted using ENMF features. The best results were achieved with codebook size of 128, sub-block overlap of 50%. ENMF
5.3. A hierarchical method

Figure 5.13: Effect of different codebook sizes (LBP feature)

Figure 5.14: Effect of different sub-block overlaps (LBP feature)
parameters were set as \( r = 121, S_h = 0.1, \alpha = 1, \psi = 0.5, \xi = 0.001 \). By employing the proposed hierarchical scheme and histogram-based representation, and compared with the results obtained by directly using ENMF coefficient (Table 3.3.4), the average recognition rate improved from 0.71 to 0.94.

**Experiments on FERET database**

The FERET face database was used to evaluate the proposed method according to the standard FERET evaluation protocol. Parameters selected for each kind of feature in the above experiments were used. In particular: codebook size of 64, sub-block overlap of 63/64, DCT subset of S5 for DCT feature; \( LBP_{(16,2)} \), codebook size of 64, sub-block overlap of 75\% for LBP feature; codebook size of 128, sub-block overlap of 50\%, \( \alpha = 1, \psi = 0.5, \xi = 0.001 \) for ENMF feature. The performance of the proposed method is shown in Table 5.2, including some reported results for the one sample problem based on the same dataset. “Proposed with predefined weight” denotes predefined weighting (4.0 for eyes regions, 3.0 for eyebrow regions, 2.0 for nose region, 1.0 for mouth region) was used for the proposed method instead of the weighting scheme based on distinctiveness. As can be seen from the table, the best results were achieved by using the proposed method with ENMF feature. By employing the proposed hierarchical scheme, and compared with the results shown in Table 5.1 with only the weighting method modified, the performance of LBP based method was further improved.

**5.4 Summary**

Motivated by the finding that distinctiveness plays an important role in human face recognition, the proposed computational model for distinctiveness measure is used to adaptively weight (select) facial regions from which local features are extracted. Based on the weighting scheme, a hierarchical method is further developed to effectively incorporate the configural information of faces for recognition. Experimental results have shown that the adaptive weighting scheme is able to substantially improve the performance of existing local-based
Table 5.2: The recognition rates of different methods for the standard FERET evaluation protocols

5.4. Summary

face recognition methods and that the proposed hierarchical method is effective.
In this chapter, the single sample face recognition problem is addressed based on the computational model for familiarity. We first introduce the concept of familiarity space, a vector space formed by the familiarity of faces measured with respect to different reference sets. Then, a novel face recognition method is proposed by analyzing the faces in the familiarity space. In particular, the proposed method considers each given face as a query to retrieve similar reference sets in the familiarity space, and examines the order in which the reference sets are retrieved. Experimental results based on the standard evaluation protocols have verified the efficacy of the proposed method.

6.1 Familiarity space

The proposed familiarity model measures the familiarity of a given face image with respect to a reference set. It actually reveals the characteristics of the face that are most common to the faces in the reference set. By employing multiple reference sets, different characteristics of the face could be exploited. Such a process of measuring the degrees of familiarity with respect to multiple reference sets is consistent with the human visual process in which we learn a face by comparing it with different previously encountered faces. From conventional pattern recognition point of view, the different reference sets act implicitly like a filter or feature extractor defined respectively by the sample faces in the reference sets. Therefore, it is conjectured that a subject could be effectively characterized by the familiarity values measured with respect to multiple and properly chosen reference sets.
6.1. Familiarity space

Assume that a large face database $T$ is divided into $q$ reference sets of faces $T^1, T^2, \ldots, T^i, \ldots, T^q$, where $T^i \subseteq T$ and, $T^i$ and $T^j$, $\forall i, j, i \neq j$, are not necessarily exclusive. The $q$ reference sets define a $q$-dimensional feature space, referred to as a familiarity space, in which each point corresponds to a face and is represented by a vector, $v_s$, of familiarity values of the face, $s$, measured against the $q$ reference sets. Thus we can write,

$$v_s = \left[ F(s, T^1), F(s, T^2), \ldots, F(s, T^i), \ldots, F(s, T^q) \right], \quad (6.1)$$

where $s$ is a face sample. Equation 6.1 can be seen as a data-dependent mapping which embeds $s$ in the $q$-dimensional familiarity space. Obviously, the discriminative power of the familiarity space for face recognition depends on the choice of $T^i$. Intuitively, each reference set $T^i, i = 1, 2, \cdots, q$ should be chosen to reveal specific facial characteristics. In addition, the more reference sets are used, the more discriminative the familiarity space could be.

6.1.1 Discriminative power of familiarity spaces

To illustrate how two subjects are discriminated in a familiarity space, the ORL face database was used as the generic dataset $T$ and was divided into 40 subsets based on the identity of the faces; each subset corresponds to one subject. Familiarity vectors of three sample faces were obtained using the method described in Section 4.2.1 by setting $C = 1$. They are shown as curves in Figure 6.1. As can be seen, face1 and face2 are images from the same person. Although the face images are with different facial expressions, their familiarity vectors are very similar. As for face3 which represents a different person from face1 and face2, its familiarity vector has a different pattern from those of face1 and face2.

To demonstrate the potential of conducting single sample-based face recognition in a familiarity space, a simple nearest neighbour (NN) classifier based on Euclidean distance was designed. The classifier, as shown in Figure 6.2, compares the familiarity vector of a test face with the familiarity vectors of the sample faces and assigns to test face the identity of the sample face that results in the minimum distance. Experiments were conducted according to the standard FERET evaluation protocols [119]. In the experiments, $T$ consisted of all face
6.2. Matching with familiarity

Like other vector spaces, different metrics other than Euclidean distance can be employed in the familiarity space. To further improve the performance of single sample-based face recognition in a familiarity space, a suitable metric should be used. However, finding such a metric is non-trivial as the underlying structure of the manifold on which faces of one
6.2. Matching with familiarity

Figure 6.2: Face recognition using a NN classifier in a familiarity space

<table>
<thead>
<tr>
<th>Methods</th>
<th>fafb</th>
<th>fafc</th>
<th>dup.I</th>
<th>dup.II</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA [149]</td>
<td>0.85</td>
<td>0.65</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td>LDA [7]</td>
<td>0.94</td>
<td>0.73</td>
<td>0.55</td>
<td>0.31</td>
</tr>
<tr>
<td>EBGM Optimal [158]</td>
<td>0.90</td>
<td>0.42</td>
<td>0.46</td>
<td>0.24</td>
</tr>
<tr>
<td>LBP [3]</td>
<td>0.97</td>
<td>0.79</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>SIFT.OCI [138]</td>
<td>0.99</td>
<td>0.92</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>Familiarity vector (NMF)</td>
<td>0.87</td>
<td>0.71</td>
<td>0.61</td>
<td>0.55</td>
</tr>
<tr>
<td>Familiarity vector (LNMF)</td>
<td>0.89</td>
<td>0.73</td>
<td>0.62</td>
<td>0.58</td>
</tr>
<tr>
<td>Familiarity vector (NMFsc)</td>
<td>0.91</td>
<td>0.69</td>
<td>0.61</td>
<td>0.53</td>
</tr>
<tr>
<td>Familiarity vector (ENMF)</td>
<td>0.94</td>
<td>0.75</td>
<td>0.68</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 6.1: The recognition rates of different methods based on the standard FERET evaluation protocols
6.2. Matching with familiarity

Figure 6.3: The conceptual diagram of the proposed method for matching two faces in a familiarity space

particular person reside in the familiarity space is almost impossible to be analyzed when only a single sample is available. Furthermore, without knowing the distributions of the faces of subjects, it can be hard to predict if two faces belong to the same person even when they are very close to each other in the space. Nevertheless, since the familiarity space is defined by the chosen multiple reference face sets, the distribution of each reference face set in the space is clearly known and can be utilized.

In this section, a novel method for matching two faces by fully leveraging the available reference sets in the familiarity space is proposed. Figure 6.3 illustrates the concept of the proposed matching method. In particular, all faces from the reference sets and the two faces to be matched are represented in the familiarity space using familiarity vectors. Each of the two faces (one is a test face and the other is the single sample of a subject) is considered as a query face to retrieve and order the reference sets. Matching decision of the two faces is then made based on their retrieval orders.

Although the proposed method relies on reference sets, it differs in many ways from the
the approaches that use generic samples as reviewed in Chapter 2. Firstly, no assumption is made about the reference faces. The reference faces can be arbitrary and may or may not contain any instances of the same subject of a test face and the (training) sample faces. Secondly, rather than directly making use of the identity information of generic samples like the similarity-based methods [37] [117] [19] or training specific classifiers like the classifier stacking-based methods [160] [5] [67], the knowledge of faces is modeled by dividing the face dataset into non-exclusive reference sets such that each reference set reveals a certain face characteristic. Furthermore, the proposed method differs from the one-shot/two-shot similarity [159], in that it does not learn models for the given samples, rather, it compares the samples by examining if they have similar neighbouring reference sets. Thirdly, the reference samples are not necessarily labeled. Finally, the proposed method is different from conventional transfer learning. A familiarity space allows the representation of a face based on the knowledge of previously encountered faces and the reference samples are not necessarily related to the two to-be-matched samples as required in any transfer learning-based methods.

6.2.1 Problem formulation

Given two samples $s_1$ and $s_2$ and a set $T$ of arbitrary samples. Assume the set $T$ can be divided into $q$ non-exclusive reference sets: $T^1, T^2, \ldots, T^q$ and consider that $s_1$ and $s_2$ are two queries to retrieve and order the $q$ reference sets. Let $O(s, T)$ denote the order of the reference sets retrieved by $s$ from the most fitted one to $s$ to the least fitted one. The matching decision of $s_1$ and $s_2$ is made as:

$$S(s_1, s_2) = D(O(s_1, T), O(s_2, T))$$ (6.2)

where $D(\cdot)$ denotes a decision making function based on the two orders. The order $O(s, T)$ is based on the fitness of the sample $s$ to the reference sets.
6.2. Matching with familiarity

6.2.2 Fitness measure

A Bayesian criterion based on marginal likelihoods [47] is adopted to measure how much $s$ fits to a reference set $T^i$. The fitness is defined as the ratio of the joint probability of observing $s$ and $T^i$ to the probability of independently observing $s$ and $T^i$.

$$fitness(s, T^i) = \frac{p(T^i, s)}{p(T^i)p(s)} \quad (6.3)$$

Intuitively, this ratio compares the probability that $s$ and $T^i$ are generated by the same model with the same parameters to the probability that $s$ and $T^i$ are realized from the models with different parameters. The larger this score, the more likely it is that the sample $s$ belongs to $T^i$. Each of the three terms in Equation 6.3 are marginal likelihoods and can be written as the integrals over a parameter space:

$$p(s) = \int p(s|\theta)p(\theta)d\theta \quad (6.4)$$

where $\theta$ are the parameters of a distribution modeling the face samples of an individual, $p(\theta)$ is the prior and $p(s|\theta)$ is the likelihood of observing $s$ given the parameter $\theta$.

For each reference subset $T^i$ that consists of $m$ face samples, the prior probability of $T^i$ can be expressed as

$$p(T^i) = \int \left[ \prod_{j=1}^{m} p(r_{ij}|\theta) \right] p(\theta) \, d\theta. \quad (6.5)$$

Here, $r_{ij}$ is the $j'th$ sample of $T^i$. Similarly, for the joint distribution of $s$ and $T^i$:

$$p(T^i, s) = \int \left[ \prod_{j=1}^{m} p(r_{ij}|\theta) \right] p(s|\theta) \, p(\theta) \, d\theta \quad (6.6)$$

It is very difficult to compute the marginal likelihoods, thus it is assumed that the models are from an exponential family, for which marginal likelihoods can be computed analytically due to the fact that they have conjugate priors. Then, the distribution $p(s|\theta)$ can be written in the form of

$$p(s|\theta) = f(s)g(\theta)\exp\left\{\theta^T u(s)\right\} \quad (6.7)$$
where \( u(s) \) is a k-dimensional vector of sufficient statistics, and \( f() \) and \( g() \) are non-negative functions describing the statistics of the data and parameters respectively.

The conjugate prior is

\[
p(\theta|\eta, \nu) = h(\eta, \nu)^{g(\theta)^\eta} \exp \left\{ \theta^T \nu \right\}
\]

where \( \eta, \nu \) are hyperparameters and \( h() \) normalizes the distribution. Then,

\[
\text{fitness}(T^i, s) = \frac{h(\eta + 1, \nu + u(s)) h(\eta + m, \nu + \sum_j u(r_{ij}))}{h(\eta, \nu) h(\eta + m + 1, \nu + u(s) + \sum_j u(r_{ij}))}
\]

For the case of Normal distribution:

\[
p(T^i | \Sigma, \mu) = \prod_{j=1}^m \frac{1}{(2\pi)^{d/2} \Sigma^{1/2}} \exp \left( -\frac{1}{2} (r_{ij} - \mu)^T \Sigma^{-1} (r_{ij} - \mu) \right)
\]

where \( d \) is the dimensionality, \( \mu \) is the mean and \( \Sigma \) is the covariance matrix. The prior distribution which is conjugate to the Normal is the Normal-Inverse Wishart distribution:

\[
p(\Sigma, \mu | b, P, k, v) = \frac{k^{d/2}}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (\mu - b)^T k\Sigma^{-1} (\mu - b) \right) \frac{|P|^{v/2} |\Sigma|^{-(v+d+1)/2} e^{-\frac{1}{2} \text{trace}(P\Sigma^{-1})}}{2^{vd/2} \pi^{(d-1)/4} \Gamma_{d}^{(v+1-x)/2}}
\]

where \( b \) is the prior mean of the mean \( \mu \), \( P \) is the prior mean of the precision matrix, \( k \) is the scaling factor on the prior precision of the mean, and \( v \) is the degree of freedom. Using Gamma function, \( \Gamma \), a generalization of the factorial function, we have:

\[
p(T^i | \mu, b, P, k, v) = (2\pi)^{-\frac{vd}{2}} \frac{k^{d/2}}{|P|^{v/2}} \frac{|P'|^{v'/2}}{\pi^{d/2} \Gamma_{d}^{(v'+1-x)/2}}
\]

where \( v' = v + m \),

\[
P' = P + \frac{km}{m + k} bb^T - \frac{1}{m + k} \left( \sum_{j=1}^m r_{ij} \right) \left( \sum_{j=1}^m r_{ij}^T \right) - \frac{k}{m + k} \left( b \left( \sum_{j=1}^m r_{ij} \right) + \left( \sum_{j=1}^m r_{ij} \right) b^T \right) + XX^T
\]

and \( X \) is the observed data.
6.2.3 Matching

The decision making function $D(\cdot)$ in Equation 6.2 is to score how much the two sample $s_1$ and $s_2$ match or to make a decision on whether they are from the same subject. For the former, a simple way is to calculate the correlation of the two orders $O(s_1, T)$ and $O(s_2, T)$. For the latter, a binary classifier can be employed.

To compute the correlation of the two orders $O(s_1, T)$ and $O(s_2, T)$, Spearman’s rank correlation (Spearman’s rho) [60] is adopted in this thesis, that is,

$$
S(s_1, s_2) = D(O(s_1, T), O(s_2, T)) = \left( \sum_i |K(s_1, T^i) - K(s_2, T^i)|^2 \right)^{1/2}.
$$

(6.14)

where $K(s, T^i)$ is the ranking of reference set $T^i$ after sorting all the subsets of $T$ based on the degrees of the fitness of $s$ to $T^i$, $i = 1, 2, \cdots, q$.

To make a binary decision on whether $s_1$ and $s_2$ have the same identity, a Support Vector Machine (SVM) [51] is adopted. The input to the SVM is the difference vector $e_{s_1, s_2}$ of the pair of orders $O(s_1, T)$ and $O(s_2, T)$:

$$
e_{s_1, s_2} = \left[ |K(s_1, T^1) - K(s_2, T^1)|, |K(s_1, T^2) - K(s_2, T^2)|, \ldots, |K(s_1, T^q) - K(s_2, T^q)| \right].
$$

(6.15)

6.2.4 Experimental results

Experiments were first conducted to evaluate the effectiveness of the fitness measurement and the proposed matching method was then evaluated using Spearman’s rank correlation and SVM for single sample-based face identification on the standard FERET database [119] and single sample-based face verification on the latest challenging LFW database [54], respectively. In all the experiments, familiarity was measured using the method described in Section 4.2.1 by setting $C = 1$.

Results on the fitness measure

The AR database was used in evaluating how effective the proposed fitness measurement is. The database contains over 4,000 color face images from 126 subjects, at least 26 face
images per subject. The 26 images are indexed from 1-26 to specify different features of the images:

- 1: Neutral expression
- 2: Smile
- 3: Anger
- 4: Scream
- 5: left light on
- 6: right light on
- 7: all side lights on
- 8: wearing sun glasses
- 9: wearing sun glasses and left light on
- 10: wearing sun glasses and right light on
- 11: wearing scarf
- 12: wearing scarf and left light on
- 13: wearing scarf and right light on
- 14-26: second session with the same conditions as 1 to 13

Face images of 60 subjects in AR database were selected and used in this experiment. All the images were first converted into gray scale and then normalized as described in Chapter 4. Figure 6.4 shows the normalized images of one subject with indexes of 1 to 13. Images of the same subject with indexes of 14 to 26 are shown in Figure 6.5. For each subject, images with indexes of 3, 6, 9, 16, 19, 22 were selected as testing query samples. Images with indexes of 1, 2, 4, 5, 7, 8, 10, 11, 12, 13 were used to form one reference set, and 14, 15, 17, 18, 20, 21, 23, 24, 25, 26 were used to form another reference set. Therefore,
there are a total of 120 reference sets and 360 testing query samples. Each subject has two corresponding reference sets. In the experiments, the rates at which each query sample face retrieves the related reference sets were recorded and summarized as follows:

- 96% top 1 retrieved reference set is with the same identity as the query sample.
- 90% top 2 retrieved reference sets are both with the same identity as the query sample.
- 92% top 3 retrieved reference sets contain the two sets that are with the same identity as the query sample.
- 97% top 4 retrieved reference sets contain the two sets that are with the same identity as the query sample.
- 99% top 5 retrieved reference sets contain the two sets that are with the same identity as the query sample.
- 100% top 6 retrieved reference sets contain the two sets that are with the same identity as the query sample.

Although the testing dataset has large variations of expressions, illuminations and occlusions, the proposed fitness measurement was effective and reference sets with the same identity as the query were always ranked in the top 6.
Table 6.2: The recognition rates of different methods for the standard FERET evaluation protocols

<table>
<thead>
<tr>
<th>Methods</th>
<th>fafb</th>
<th>fafc</th>
<th>dup.I</th>
<th>dup.II</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA [149]</td>
<td>0.85</td>
<td>0.65</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td>LDA [7]</td>
<td>0.94</td>
<td>0.73</td>
<td>0.55</td>
<td>0.31</td>
</tr>
<tr>
<td>EBGM Optimal [158]</td>
<td>0.90</td>
<td>0.42</td>
<td>0.46</td>
<td>0.24</td>
</tr>
<tr>
<td>LBP [3]</td>
<td>0.97</td>
<td>0.79</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>SIFT_OCI [138]</td>
<td>0.99</td>
<td>0.92</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td>Familiarity vector + Euclidean distance</td>
<td>0.94</td>
<td>0.75</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>Proposed method using correlation</td>
<td>0.98</td>
<td>0.89</td>
<td>0.77</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Results of identification

For single sample-based face identification, the standard FERET evaluation protocols [119] were used. During the test, the same dataset \( T \) as the one used for testing the simple classifier in Section 6.1.1 was created. It consists of all the face images from ORL database and face images of 60 subjects from the AR database. K-means was employed to cluster \( T \) into different \( n \) numbers of exclusive groups, where \( n = 5, 10, 20, 50, 100 \). All the clustered groups obtained were used as reference sets. Thus 185 non-exclusive reference sets in total were generated. In the experiment, the test face was decided to belong to the subject whose sample face gives the maximum value of the Spearman’s correlation (Equation 6.14). The results are shown in Table 6.2. For the convenience of comparison, the results obtained by other popular recognition methods and the simple classifier based on familiarity vectors are also presented in the Table 6.2. As seen, the proposed matching method has significantly improved the performance for all protocols compared to the simple classifier based on familiarity vectors and has also substantially improved the performance for the protocols dup.I and dup.II compared to the SIFT_OCI method [138].

Results of verification

Experiments of face verification were conducted on the LFW-a version of view-2 of the LFW database [54]. This part of the LFW was produced by aligning all the original LFW images using the commercial alignment system, face.com and provides a benchmark protocol called
“image-restricted training”. In the protocol, the benchmark data consists of 6000 pairs of faces, half marked “same” and half not, and is divided into 10 equally sized sets and the benchmark test was repeated 10 times, each time using one set for testing and nine others for training. The goal is to predict which of the test pairs match using only the specified training data. Other than “same/not-same” labels, no information is provided on the identity of the subjects.

During the test on this benchmark, SVM was used to make the decision. One of the 9 training splits was employed as the dataset $T$, the other 8 were used to train a SVM classifier for “same or not same” with an RBF kernel. $T$ was first clustered into different number $n$ of exclusive groups, with $n = 5, 10, 20, 50, 100$, and then all these clustered groups were used as the non-exclusive reference sets. The test results together with reported verification accuracies of other recent works on the same dataset are shown in Table 6.3. We can see that the proposed method outperformed most of the state-of-the-art methods and achieved comparable performance to the work of Kumar et al. [67]. However, its performance is not as good as that of the Associate-Predict method [175]. This is likely due to the fact that both works, [67] and [175], utilized extra labeled training data in addition to the LFW data. Particularly, the “attribute classifiers” and “simile classifiers” used in [67] were trained based on a huge volume of labeled images outside the LFW dataset. The Associate-Predict method [175] requires reliable facial feature point detectors and employed a subset of the well labeled Multi-PIE dataset (covering 7 poses and 4 illumination conditions per subject) as generic samples.

6.3 Summary

This chapter has introduced the concept of a familiarity space defined by multiple reference sets and studied how to represent and compare faces in the space. In particular, a novel single sample-based face recognition method has been developed. The results on the popular benchmarking datasets have demonstrated the discriminative power that a familiarity space could potentially offer. Studies on the choice of reference sets to form a discriminative
Table 6.3: The verification accuracies of different methods based on the image-restricted training protocol in LFW database view 2

<table>
<thead>
<tr>
<th>Method</th>
<th>Verification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDML[91]</td>
<td>0.793</td>
</tr>
<tr>
<td>POEM[153]</td>
<td>0.754</td>
</tr>
<tr>
<td>DML-eig SIFT[176]</td>
<td>0.813</td>
</tr>
<tr>
<td>Single LE[174]</td>
<td>0.812</td>
</tr>
<tr>
<td>One-shot similarity[159]</td>
<td>0.820</td>
</tr>
<tr>
<td>Two-shot similarity[159]</td>
<td>0.659</td>
</tr>
<tr>
<td>Attribute and Simile classifiers[67]</td>
<td>0.837</td>
</tr>
<tr>
<td>Associate-Predict [175]</td>
<td>0.906</td>
</tr>
<tr>
<td>Proposed method using SVM</td>
<td>0.828</td>
</tr>
</tbody>
</table>

familiarity space and extension of the representation and recognition to other objects can form the basis of future work.
Chapter 7

Age and head pose estimation

In this chapter, the representative power of the ENMF is further exploited in the context of classification where face images in the reference set are labeled exclusively. A naive classifier is devised and evaluated in the tasks of head pose estimation and age group classification. In addition, the naive classifier is extended to a coarse-to-fine hierarchical classification scheme for exact age estimation. Results obtained from commonly used benchmarking datasets demonstrated that both the naive classifier and the hierarchical scheme outperform the state-of-the-art methods for age and head pose estimation.

7.1 A naive classifier based on ENMF representation

Assume that a reference face set $R$ consists of $C$ exclusive and labeled reference groups and that images in the same group share the same label. Let $V_i$ represents the face images in group $R_i$ ($i \in \{1, 2, \ldots, C\}$), then an ENMF subspace $W_i$ learned from $V_i$ can be regarded as a specific feature space for the group $R_i$. Given a new face image $s$, its coefficient vector $h_i$ in the group subspace $W_i$ can be written as

$$h_i = W_i^{-1}s, \quad (7.1)$$

where $W_i^{-1}$ is the pseudo inverse matrix of $W_i$. Based on the obtained coefficient vector $h_i$, the sample $s$ can be reconstructed as

$$\hat{s} = W_i h_i. \quad (7.2)$$
The reconstruction error $\epsilon_t$ between $s$ and $\hat{s}^t$ reflects the similarity between the new face and the images in the group $R_t$. From classification point of view, a smaller value of $\epsilon_t$ indicates a higher probability that $s$ belongs to the group $R_t$. Thus, after ENMF subspaces are learned for each of the $C$ groups, $s$ can be classified as belonging to the $q'th$ group, if

$$\epsilon_q = \min_{i=1,...,C} \{\epsilon_i\}$$

(7.3)

where the reconstruction error is represented as the mean square error (MSE). Hence we have

$$\epsilon_t = \text{MSE}(s, \hat{s}^t) = \frac{1}{n} \sum_i (s_i - \hat{s}_i)^2,$$

(7.4)

where $n$ is the number of pixel in the face image.

Note that the classifier (7.4) employs different feature spaces, (i.e. ENMF subspace learned from each individual group) for different groups to effectively explore the intra- and inter-group variations. This is different from most conventional classifiers where the test sample is compared with all training samples in the same single feature space. In addition, the reconstruction error $\epsilon_t$ is a special case of the familiarity measure for sample $s$ with respect to $R_t$ by setting $C = 1$ in (4.4).

In the following sections, the classifier is employed for head pose estimation, age group classification and exact age estimation and compared to the state-of-the-art methods for these face analysis tasks.

7.2 Head pose estimation

As reviewed in Chapter 2.3, although the regression-based methods have performed favourably and are able to estimate head pose at the resolution of degrees. However, these methods require a large number of training data covering all the head poses with accurate labels. In the absence of such large and accurate training sets, most of existing methods formulate head estimation as a pattern classification problem based on publicly available face datasets. Recently, Hu et al. [53] evaluated the performance of major subspace learning methods for head pose estimation based on CMU PIE face database [134]. In this section, we employ the
7.2. Head pose estimation

Figure 7.1: Sample face images of one person from CMU PIE database; the numbers beneath each image represent the poses

naive classifier to estimate head poses where the reference face set consists of labeled groups of faces representing different poses, and compare the naive classifier with those methods evaluated in [53] using similar experimental set-up.

The CMU PIE face database contains 41,368 face images of 68 subjects, the images were captured by 13 synchronized cameras (thus under 13 different poses) under varying illumination and expressions. A subset of the database was used in the experiment: for each of the head pose, 408 images were selected, 6 images per subject, with expression and illumination variations (same set of images as used in [53]). Half of the images (from the first 34 subjects) were used for learning 13 pose subspaces and the rest of data were used for testing. Rather than cropping and aligning the face areas based on key facial points as in Hu et al.’s work, the face regions of the images were detected by the Viola-Jones face detection method [152] and cropped without alignment. In the few images where the faces were not detected automatically, the face areas were manually cropped. All the cropped face images were then converted to gray scale images and resized to 64 × 64. Some sample images are shown in Figure 7.1.

While fixing the number of bases \( r \) to 49, the proposed classifier was first tested with different value of \( S_h \) (the sparseness of coefficient matrix). The results are shown in Figure 7.2. We can see that the best result was achieved neither with the highest nor with the lowest sparseness, but when \( S_h \) is set to 0.3. This observation justifies our analysis in Chapter 3,
7.2. Head pose estimation

Figure 7.2: The testing results for different value of $S_h$ while $r$ is fixed to 49 and demonstrates that a compromise made between localization and overlapping improves the efficiency of NMF for head pose representation.

Then, we set the value of $S_h$ to 0.3 and tested the classifier by changing the number of bases, $r$. The results are shown in Figure 7.3. As can be seen from the figure, high estimation accuracy were obtained as the number of bases increases. However, the estimate accuracy was almost saturated when the value of $r$ exceeds 81.

Table 7.1 lists the best result we obtained along with the accuracies of other major subspace-based head pose estimation methods reported in [53]. The estimation accuracy for each pose are shown in Figure 7.4. Hu et al. evaluated four methods in their work: Eigenface method (PCA), Fisherface method (LDA), locality preserving projections (LPP), and Pose Specific Subspace (PSS). In the first three methods, a given face is projected onto one subspace for all the poses, the projected face is then compared to the stored pose representation and classified to one of the pose categories. PCA seeks a projection that best represents the data in a least-squares sense, LDA looks for directions that are efficient for discrimination, while LPP tries to find an embedding that preserves intrinsic geometry of the data and local information. Similarly to the proposed method, PSS models each pose with one individual eigenspace, the distance from the pose specific subspace is then exploited as the similarity measurement for pose estimation. We can see from Figure 7.4 that the proposed method almost outperforms the other methods for all the head poses, especially for pose 27 (frontal pose), which is relatively hard to classify correctly since there are four near frontal views. As can be seen from Table 7.1, LPP and PSS performed better than the traditional subspace methods - PCA and LDA. By focusing on the local information and characterizing different poses in their own subspaces, the proposed method combines the advantages of both LPP
7.3 Age group classification

In the second application, the naive classifier is employed to classify a face into one of the predefined age groups. The reference set consists of face images that are labeled exclusively into different groups of ages.

The FG-NET [1] and MORPH [125] databases were used for performance evaluation and

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of basis</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>117</td>
<td>75.72%</td>
</tr>
<tr>
<td>LDA</td>
<td>13</td>
<td>75.37%</td>
</tr>
<tr>
<td>LPP</td>
<td>87</td>
<td>78.90%</td>
</tr>
<tr>
<td>PSS</td>
<td>80</td>
<td>83.19%</td>
</tr>
<tr>
<td>Proposed</td>
<td>121</td>
<td>87.82%</td>
</tr>
</tbody>
</table>

Table 7.1: The average accuracies of subspace-based head pose estimation methods and PSS, and thus achieved the best result.

Figure 7.3: The testing results for different value of $r$ while $S_h$ is fixed to 0.3

Figure 7.4: Comparison of the estimation accuracy for each pose by different methods
7.3. Age group classification

<table>
<thead>
<tr>
<th>Age group</th>
<th>MORPH (Album 2)</th>
<th>FGNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>0%</td>
<td>37%</td>
</tr>
<tr>
<td>10-19</td>
<td>13%</td>
<td>33%</td>
</tr>
<tr>
<td>20-29</td>
<td>29%</td>
<td>14%</td>
</tr>
<tr>
<td>30-39</td>
<td>28%</td>
<td>8%</td>
</tr>
<tr>
<td>40-49</td>
<td>22%</td>
<td>5%</td>
</tr>
<tr>
<td>50-59</td>
<td>7%</td>
<td>2%</td>
</tr>
<tr>
<td>60-69</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>

Table 7.2: Age group distribution of the images in MORPH and FG-NET database

The FG-NET database contains 1002 face images of 82 multiple-race subjects with large variations in lighting, pose, and expression. For each image, the database also provides the locations of manually annotated 68 facial landmarks. The MORPH database consists of two albums of face images. Only the second album was used in our experiments. This album contains 55,134 images of 13,000 individuals collected over four years with applicable metadata on race, gender, date of birth, and date of acquisition. Table 7.2 shows age group distributions of the images in the two databases. As can be seen in the table, the age distributions of both databases are highly uneven. The second album of MORPH database does not contain face images younger than 16 and very limited face images are available for ages older than 50 in both databases.

The first experiment was conducted to tune the two ENMF parameters, \( r \) and \( S_h \), using the same process as in the head pose estimation. We considered the ages ranging from 0 to 44 and divided them into 5 groups: 0 – 5 (preschool age), 6 – 11 (primary school age), 12 – 17 (secondary school age), 18 – 29 (young adult) and 30 – 44 (middle age). Images from 50 randomly selected subjects were used as training set and images for the rest 32 subjects were all used for testing. For each age group, 90 images were randomly selected from the training set to learn the corresponding age group subspace. Figure 7.5 shows the average classification accuracy versus the value of \( S_h \) while fixing \( r \) at 81. The best accuracy was obtained at \( S_h = 0.3 \). Figure 7.6 shows the classification accuracy versus the number of bases \( r \), which indicates that when \( r \geq 121 \), the classification accuracy almost reaches the best. Therefore, we set \( r \) to 121 and \( S_h \) to 0.3 for the rest of the experiments.

In the second experiment we compare the performance of the naive classifier with that
7.3. Age group classification

Figure 7.5: The classification accuracy for different values of $S_h$ while $r = 81$

Figure 7.6: The classification accuracy for different values of $r$ while $S_h = 0.3$

of Kriegel’s work [66] which employs AAM features and SVM classifiers. In [66], the FG-NET database was divided into 4 age groups in the ranges of $0 – 19$, $20 – 29$, $30 – 39$ and $40 – 69$ for the age group classification. Kriegel [66] also conducted 2 age group classification by separating the FG-NET database at the ages of 10, 14, 18, 20 and 30. Under the same experimental set-up, our naive classifier was employed to conduct the age group classification. The results along with classification accuracies of the AAM+SVM method are listed in Table 7.3 and Table 7.4. As can be seen from the tables, the naive classifier achieved better results in all the cases especially for the young age groups. This is mainly because the naive classifier emphasizes the local information and characterize each age group in its own subspace.

The third experiment was to compare the naive classifier with the AGES, AAS and WAS methods. Geng et al. [35] evaluated the performance of their AGES age estimator as well as Lanitis et al.’s AAS [71] and WAS [72] methods in age group classification on the FG-NET database. Following the same experimental set-up as that in [35], ages from 0 to 69 were
### 7.3. Age group classification

<table>
<thead>
<tr>
<th>Age Group</th>
<th>AAM+SVM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-19</td>
<td>71.7%</td>
<td>84.6%</td>
</tr>
<tr>
<td>20-29</td>
<td>40.7%</td>
<td>49.7%</td>
</tr>
<tr>
<td>30-39</td>
<td>45.8%</td>
<td>51.1%</td>
</tr>
<tr>
<td>40-69</td>
<td>50.0%</td>
<td>58.2%</td>
</tr>
<tr>
<td>Overall</td>
<td>52.1%</td>
<td>60.9%</td>
</tr>
</tbody>
</table>

Table 7.3: The classification accuracy for 4 age groups

<table>
<thead>
<tr>
<th>Age Threshold</th>
<th>AAM+SVM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>75.9%</td>
<td>87.1%</td>
</tr>
<tr>
<td>14</td>
<td>73.2%</td>
<td>85.9%</td>
</tr>
<tr>
<td>18</td>
<td>72.2%</td>
<td>85.3%</td>
</tr>
<tr>
<td>20</td>
<td>71.5%</td>
<td>84.0%</td>
</tr>
<tr>
<td>30</td>
<td>76.0%</td>
<td>86.3%</td>
</tr>
</tbody>
</table>

Table 7.4: The classification accuracies for 2 age groups

Divided into 14 age groups, each of which covers 5 continuous ages, i.e. $0-4, 5-9, \ldots, 65-69$. Since the number of face images for ages older than 44 in FG-NET is too small to learn ENMF subspaces, these images were all used for testing. Instead, the ENMF subspaces were trained using images from MORPH database for age groups in the range 45-69. As for ages of 0-44, all training and testing images were from the FG-NET database and four-fold cross validation was performed. The average classification accuracy of the naive classifier as well as the accuracies of AGES, WAS and AAS reported in [35] are shown in Table 7.5. To further illustrate the efficiency of the proposed ENMF representation, traditional NMF, LNMF and NMFsc were also used to learn age group subspaces for classification. The results are included in Table 7.5. As can be seen in the table, the ENMF based naive classifier achieved the best classification accuracy and the ENMF has clear advantages over NMF, LNMF and NMFsc. It has to be pointed out that the four-fold cross validation and cross-database validation for the groups with ages between 45 and 69 represent a much more rigorous evaluation than the Leave-One-Person-Out (LOPO) strategy used in [35] for evaluating the AGES, AAS and WAS methods.
7.4 Exact age estimation

In this section, we extend the age group classification for exact age estimation. Specifically, a coarse-to-fine hierarchical classification scheme [71] is constructed using the naive classifier. However, a common problem with the hierarchical scheme is that the errors occurring at the top-level can propagate to the sub-level since border ages tend to be misclassified into neighbor groups. To reduce the error propagation, additional border age groups that sit around the boundary of two regularly partitioned adjoining age groups are introduced during the top-level age group classification. The sub-level age estimation is conducted within one selected regular age group as well as one selected border group (as shown in Figure 7.7). In particular, after ENMF subspaces are learned for each age group and each specific age, the exact age of a given face image \( s \) is estimated through the following steps:

1. Project \( s \) onto each of the regular age group subspaces; select the age group with minimum reconstruction error.

2. Project \( s \) onto each of the border age group subspaces; select the border age group with minimum reconstruction error.

3. Project \( s \) onto each of the specific age subspaces within the range of the ages defined by the selected age group and border age group; select the age with the minimum reconstruction error.

The performance of the proposed coarse-to-fine age estimator was evaluated using the FG-NET and MORPH databases. Due to insufficient number of images in the databases for

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGES</td>
<td>40.92%</td>
</tr>
<tr>
<td>WAS</td>
<td>18.16%</td>
</tr>
<tr>
<td>AAS</td>
<td>26.65%</td>
</tr>
<tr>
<td>NMF</td>
<td>42.37%</td>
</tr>
<tr>
<td>LNMF</td>
<td>44.16%</td>
</tr>
<tr>
<td>NMFsc</td>
<td>41.24%</td>
</tr>
<tr>
<td>ENMF</td>
<td>45.77%</td>
</tr>
</tbody>
</table>

Table 7.5: The classification accuracy for 14 age group estimations
some ages, we only considered the ages ranging from 15 to 59. For each age from 15 to 59, 90 face images (mainly faces of Caucasians) were selected from MORPH database Album 2 as training data to learn the ENMF subspaces of different specific ages and age groups. For the coarse age group classification, 5 regular age groups (15-23, 24-32, 33-41, 42-50, 51-59), and 4 border age groups (21-26, 30-35, 39-44, 48-53) were used. All images within the considered age range from FG-NET database were used as testing samples. The performance was measured by the mean absolute error (MAE), which is defined as the average of the absolute errors between the estimated ages and the ground truth. Results were compared with the reported results of the state-of-the-art methods for age estimation.

Table 7.6 shows the MAEs of different methods for different age ranges. MAEs of other works are the reported results on FG-NET database using LOPO strategy. As can be seen from the table, the proposed method achieved the best MAE (the ENMF column) for age range 10-19 and 50-59. For other ages, our method is comparable and even superior to the other methods in many cases.

To see the contribution of the border age groups, the method was also evaluated by only using regular age groups during the age group classification and results are listed in the ENMF(nb) column. By using additional border age groups at the top level age group classification to reduce the error propagation, we can see that the performance of the proposed method has been highly improved.
### 7.5 Summary

A naive classifier is devised based on ENMF representation to utilize the class labeling information in the reference set. Individual ENMF subspaces are learned for the classes. The class of a given sample is then estimated based on its reconstruction errors after being projected into the learned subspaces. Experiments of cross validation and cross-database validation based on commonly used benchmarking datasets have showed that the naive classifier performs well in the tasks of head pose and age estimation.

<table>
<thead>
<tr>
<th>Age</th>
<th>ENMF</th>
<th>ENMF(nb)</th>
<th>BIF [41]</th>
<th>AGES [34]</th>
<th>RUN [165]</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-19</td>
<td>3.21</td>
<td>3.59</td>
<td>3.39</td>
<td>3.83</td>
<td>3.76</td>
</tr>
<tr>
<td>20-29</td>
<td>7.46</td>
<td>8.15</td>
<td>4.30</td>
<td>8.01</td>
<td>6.38</td>
</tr>
<tr>
<td>30-39</td>
<td>13.23</td>
<td>17.46</td>
<td>8.24</td>
<td>17.91</td>
<td>12.51</td>
</tr>
<tr>
<td>40-49</td>
<td>17.17</td>
<td>25.96</td>
<td>14.98</td>
<td>25.26</td>
<td>20.09</td>
</tr>
<tr>
<td>50-59</td>
<td>18.83</td>
<td>25.21</td>
<td>20.49</td>
<td>36.40</td>
<td>28.07</td>
</tr>
</tbody>
</table>

Table 7.6: The MAE for exact age estimation, ENMF(nb) denotes the proposed method without using border groups
Chapter 8

Conclusion

8.1 Summary

Results from extensive psychological studies indicate that humans make effective use of previously encountered faces in forming their internal face representation and driving the process of face perception. This thesis proposes to use reference samples that may or may not contain any labeling information and any instances of the same sample of the face under consideration to model previously encountered faces. Computational models have been established to account for the key process involved in face perception.

Specifically, to imitate the psychological abstract feature model advocated by Wallis et al. [155], non-negative matrix factorization (NMF) has been extended and used in learning compact localized representation of face images. The extension imposed an orthogonality constraint on the basis matrix and controlled the sparseness of the coefficient matrix. The proposed extended NMF (ENMF) overcomes the problems of the conventional NMF and its major variations that are sensitive to the variations and misalignment of the training samples.

Based on the ENMF face representation, computational models have been developed for the two fundamental processes involved in the face perception: locating distinctive areas and measuring the familiarity of a face. The distinctiveness model measures the degree of distinctiveness of any facial area based on the probability of this area belonging to a general face in the ENMF subspace learned from the reference samples. The familiarity model provides computational basis for the context-free and context-dependent forms of familiarity and accounts for the key factors: exposure frequency, exposure intensity, similar
8.2 Future works

The computational models for distinctiveness and familiarity have been further verified by applying them to single sample-based face recognition and estimation of head pose and ages. The distinctiveness model is used to adaptively weight (select) facial regions from which local features are extracted. A hierarchical method has been proposed to effectively incorporate the configural information with the distinctiveness for single sample-based face recognition. With the familiarity model, a concept of familiarity space was introduced and preliminary study has been conducted to match two faces in the familiarity space. Experiments have shown that the proposed models together with the flexibility of organizing/choosing the reference samples have offered a new paradigm and effective tools to improve the existing facial analysis methods and/or develop novel methods.

8.2 Future works

The computational models for face distinctiveness and familiarity, the concept of familiarity spaces and the reference-based face recognition developed in this thesis have provided a set of useful tools not only for computational facial analysis but also for further psychological study of human visual perception. In addition, the thesis has established a new approach for objection recognition using reference samples and, at the same time, has revealed many interesting research problems.

- The proposed models for face representation, familiarity and distinctiveness can potentially be extended and integrated into a unified computational model that is able to account for the whole comparative process of human face perception. Such an extension can also take into account other key factors that affect human face processing such as typicality and other-race effect [150, 109].
- As the discriminative ability of a familiarity space depends on the choice of each reference set, it is expected that the performance of the proposed method in a familiarity exposure and prototype effect. Both computational models have produced results that are highly consistent with subjective judgments.

The computational models for distinctiveness and familiarity have been further verified by applying them to single sample-based face recognition and estimation of head pose and ages. The distinctiveness model is used to adaptively weight (select) facial regions from which local features are extracted. A hierarchical method has been proposed to effectively incorporate the configural information with the distinctiveness for single sample-based face recognition. With the familiarity model, a concept of familiarity space was introduced and preliminary study has been conducted to match two faces in the familiarity space. Experiments have shown that the proposed models together with the flexibility of organizing/choosing the reference samples have offered a new paradigm and effective tools to improve the existing facial analysis methods and/or develop novel methods.

8.2 Future works

The computational models for face distinctiveness and familiarity, the concept of familiarity spaces and the reference-based face recognition developed in this thesis have provided a set of useful tools not only for computational facial analysis but also for further psychological study of human visual perception. In addition, the thesis has established a new approach for objection recognition using reference samples and, at the same time, has revealed many interesting research problems.

- The proposed models for face representation, familiarity and distinctiveness can potentially be extended and integrated into a unified computational model that is able to account for the whole comparative process of human face perception. Such an extension can also take into account other key factors that affect human face processing such as typicality and other-race effect [150, 109].
- As the discriminative ability of a familiarity space depends on the choice of each reference set, it is expected that the performance of the proposed method in a familiarity exposure and prototype effect. Both computational models have produced results that are highly consistent with subjective judgments.
space would be improved by optimizing the reference sets. The reference sets could be manually designed and selected when label information is available, or could be optimized by employing more complicated clustering techniques rather than k-means and by using multiple features. Therefore, it is interesting to further explore the ability of the proposed “familiarity-based face matching” to deal with illumination and viewpoint variations by using appropriate reference sets.

- Humans perceive faces in a similar manner to which they perceive objects. The models and concepts developed in this thesis for faces can be extended to the recognition and classification of generic objects. Such study can potentially lead to a new paradigm of object recognition.
Bibliography


